DOCUMENT RESUME

ED 167 076 IR 006 739

| | |
|---|---|
| AUTHOR | Arms, William Y.; Arms, Caroline |
| TITLE | The Application of Clustering Techniques to Citation Data. Research Reports Series B No. 6. |
| INSTITUTION | Bath Univ. of Technology (England). Univ. Library. |
| PUB DATE | Oct 77 |
| NOTE | 112p.; Design of Information Systems in the Social Sciences Series ; Parts may be marginally legible due to print quality |
| EDRS PRICE | MF-$0.83 HC-$6.01 Plus Postage. |
| DESCRIPTORS | Algorithms; *Bibliographic Citations; *Cluster Analysis; *Information Systems; Literature Reviews; Periodicals; Pilot Projects; Research Methodology; *Social Sciences |
| IDENTIFIERS | *Design of Information Systems in Social Sciences |

ABSTRACT

This report describes research carried out as part of the Design of Information Systems in the Social Sciences (DISISS) project. Cluster analysis techniques were applied to a machine readable file of bibliographic data in the form of cited journal titles in order to identify groupings which could be used to structure bibliographic files. Practical applications of the data obtained from clusters of journal title citations were to include the planning patterns of journal coverage for secondary services, ascertaining information on the structure of the primary literature of a discipline, and the provision of data to help validate classification schemes and thesauri. The analyses were carried out on a file of over 60,000 citations taken from social science journals, and a special cluster analysis was made of citations taken from criminology source journals at three different time intervals. Included here are a review of clustering techniques applicable to bibliographic data, a detailed discussion of the results, and an evaluation of the practical value of the technique. It was concluded that the use of cluster analysis is probably not a very practical aid to the design of secondary services in the social sciences, though the techniques nevertheless have important bibliographic applications elsewhere. A bibliography is appended. (Author/JD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Bath University Library

ED167076

# Design of Information Systems in the Social Sciences

Research Reports Series B no. 6

# The Application of Clustering Techniques to Citation Data

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J.H. Lamble

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

IR006739

Bath University
October 1977

PREFACE

The research project Design of Information Systems in the Social Sciences (DISISS) was carried out between 1971 and 1975 with the support of a grant from the British Library Research and Development Department (formerly the Office for Scientific and Technical Information). The central team was based at the University of Bath, assisted by staff at the Polytechnic of North London School of Librarianship and the Open University. The results of the research are reported in two series of papers. These reports can be obtained individually on loan from the British Library Lending Division, Boston Spa ; a limited number are available for purchase from the Secretary, the Library, Bath University, Claverton Down, BATH BA2 7AY.

This report describes work on the clustering of journal citations. The main work was carried out by W.Y. Arms and Mrs. C.R. Arms (Open University). The data files were obtained partly from the Institute for Scientific Information (ISI) and partly from data collected by the team, in particular D. Nicholas and Miss M. Ritchie. The data conversion was carried out by R.G. Bradshaw.

Some of this work was previously reported in DISISS Working Paper No.11, Clustering of journal titles according to citation data : report on preparatory work, design, data collection and preliminary analyses, (September 1973).

## LIST OF DISISS PUBLICATIONS

Working papers

1. Comparison of results of science user studies with "Investigation into information requirements of the social sciences". Microfiche only, 0.25p.
ISBN 900843 28 4

2. A machine readable data base of social science serials. Microfiche only, 0.25p.
ISBN 900843 29 2

3. The relationship between primary and secondary literature in the social sciences: a study of secondary literature in criminology. Microfiche only, 0.25p.
ISBN 900843 30 6

4. Characteristics of citation in social science monographs. Microfiche only, 0.25p.
ISBN 900843 31 4

5. Citation patterns in the social sciences: results of pilot citation study and selection of source journals for main citation study. Microfiche only, 0.50p.
ISBN 900843 32 2

6. The use of citation linkages and networks for information retrieval in the social sciences. Microfiche only, 0.50p.
ISBN 900843 39 X

7. Size and growth of monograph literature, with particular reference to the social sciences. Microfiche only, £1.00.
ISBN 900843 40 3

8. CLOSSS: a machine readable data base of social science serials: progress report 1971-72. Paper £1.00, microfiche 0.50p.
ISBN 900843 38 1

9. (Not used)

10. (Not used)

11. Clustering of journal titles according to citation data: preliminary report on preparatory work, design and data collection. Microfiche only, 0.50p.
ISBN 900843 43 8

Research reports

A1. Towards the improvement of social science information systems: overview of research carried out 1970-1975 (in preparation).

A2. Size, growth and structure of social science literature. March 1975. £2.50.
ISBN 900843 527

A3. The structure of social science literature as shown by citations (in preparation).

A4. The evaluation of operational effectiveness and its use in the design of information systems. February 1975. £1.50.
ISBN 900843 533

A5. The planning of indexing and abstracting services in the social sciences: coverage, overlap and content, 1976. £2.50.
ISBN 900843 551

B1. User evaluation of an information service in social welfare: report of an interview and questionnaire study of the Wiltshire Social Services Staff Digest. November 1973. £1.00; microfiche 0.25p.
ISBN 900843 454

B2. Analysis of requests made to the National Children's Bureau question and answer service. November 1973. £1.00; microfiche 0.25p.
ISBN 900843 462

B3. The relationship between usefulness and style of secondary publication. November 1973. £2.00; microfiche 0.50p.
ISBN 900843 470

B4. Characteristics of social science serials: the construction and analysis of a file of social science serial titles, 1975. £3.00.
ISBN 900843 489

B5. The collection and construction of a file of citation data, 1976. £3.00
ISBN 900843 497

B6. The application of clustering techniques to citation data, 1977
ISBN 900843 500

B7. Literature usage and interrelationship in the social sciences as shown by citations in monographs. (In press).
ISBN 900843 578

## DISISS Research Personnel

| | |
|---|---|
| PROJECT HEAD | Maurice B. Line, Director General, British Library Lending Division. |
| **UNIVERSITY OF BATH** | |
| Senior Research Fellow | J. Michael Brittain[1] |
| Research Fellow | Stephen A. Roberts[2] |
| Research Assistants | Barbara Skelton (until September 1973) |
| | Peter Burridge (from November 1973) |
| Programmer/Analyst | Robert G. Bradshaw (from February 1972)[3] |
| **THE POLYTECHNIC OF NORTH LONDON** | |
| Research Fellow | David Nicholas |
| Research Fellow | Maureen Ritchie |
| **OPEN UNIVERSITY** | |
| Consultant Systems Analyst | William Y. Arms (Part-time) |
| Programmer/Analyst | Caroline Arms (Part-time) |

---

1 Now Senior Lecturer, Department of Library and Information Studies, Loughborough University.

2 Now Research Fellow, Library Management Research Unit, Department of Library and Information Studies, Loughborough University.

3 Now Systems Programmer, Computing Centre, University College, Cardiff, Wales

CONTENTS


| | | Page |
|---|---|---|
| 1.0 | INTRODUCTION | 1 |
| | 1.1 Statement of aims | 1 |
| | 1.2 Reasons for using citation data | 2 |
| 2.0 | THE USE OF CLUSTER TECHNIQUES | 4 |
| | 2.1 Survey of previous work | 4 |
| | 2.2 Clustering terms | 6 |
| | 2.3 Clustering techniques | 6 |
| | 2.3.1 The SCICON algorithm | 8 |
| | 2.3.2 Reasons for selection of the SCICON algorithm | 10 |
| | 2.4 Comparison with other approaches and algorithms | 12 |
| 3.0 | DATA FILES | 24 |
| | 3.1 Files used | 24 |
| | 3.1.1 The pilot study data | 26 |
| | 3.1.2 Citations from criminology journals | 26 |
| | 3.1.3 Citations from the ranked and random lists of sources | 30 |
| | 3.2 Treatment of data before clustering | 35 |
| | 3.2.1 Reduction in the size of the data matrix | 35 |
| | 3.2.2 Normalization, self-citation adjustment and scaling | 36 |
| | 3.2.3 Choice of data adjustments | 38 |
| 4.0 | THE PILOT STUDY | 39 |
| | 4.1 Results | 39 |
| | 4.1.1 Unmodified data | 39 |
| | 4.1.2 With cells divided by row and column totals | 40 |
| | 4.1.3 With cells divided by row totals only | 41 |
| | 4.1.4 With cells divided by row totals and a constant added to each cell | 47 |

CONTENTS (cont'd)


| | | Page |
|---|---|---|
| | 4.1.5 Adjustments to compensate for self-citation | 48 |
| | 4.2 Results using other approaches | 48 |
| | 4.3 Conclusions from the pilot study | 52 |
| 5.0 | THE CRIMINOLOGY CLUSTERS | 53 |
| | 5.1 1950 sample | 53 |
| | 5.2 1960 sample | 56 |
| | 5.3 1970 sample | 60 |
| 6.0 | CLUSTERS FROM THE MAIN CITATION FILE | 69 |
| | 6.1 Citations from the ranked list of journals | 69 |
| | 6.2 Citations from the random list of journals | 78 |
| | 6.3 Comparison between the clusters derived from the ranked and random samples | 85 |
| 7.0 | DISCUSSION OF RESULTS | 90 |
| | 7.1 Reliability and stability of clusters | 90 |
| | 7.1.1 Date of source journals | 90 |
| | 7.1.2 Choice of sources | 90 |
| | 7.1.3 Clustering methods | 92 |
| | 7.2 Evaluation of clusters | 93 |
| | 7.2.1 Representation of clusters | 93 |
| | 7.3 The use of cluster analysis as a practical tool | 94 |
| Appendix I | | 95 |
| References | | 98 |

## TABLES


| | | Page |
|---|---|---|
| 1 | Distance functions suggested for hierarchical clustering .. .. | 14 |
| 2 | Source journals for the pilot study .. .. .. .. .. .. | 27 |
| 3 | Literature from criminology journals to other journals .. .. .. | 29 |
| 4 | Ranked list of source journals .. .. .. .. .. .. .. | 31 |
| 5 | Random list of source journals .. .. .. .. .. .. .. | 33 |
| 6 | Clusters derived from pilot data .. .. .. .. .. .. .. | 42 |
| 7 | Cluster characteristics and descriptions .. .. .. .. .. | 46 |
| 8 | Number of criminology journals clustered .. .. .. .. .. | 53 |
| 9 | Criminology journals in 1960 and 1970 clusters .. .. .. .. | 67 |
| 10 | Overlap between ranked and random samples of journals .. .. .. | 69 |
| 11 | Thirty clusters from ranked sample .. .. .. .. .. .. .. | 70 |
| 12 | Possible labels for thirty clusters from the ranked sample .. .. | 75 |
| 13 | Ranked sample - number of non-hierarchical journals .. .. .. | 77 |
| 14 | Thirty clusters from random sample .. .. .. .. .. .. .. | 79 |
| 15 | Possible labels for thirty clusters from the random sample .. .. | 82 |
| 16 | Random sample - number of non-hierarchical journals .. .. .. | 84 |
| 17 | Subjects of journals in the ranked sample .. .. .. .. .. | 86 |
| 18 | Subjects of journals in the random sample .. .. .. .. .. | 87 |
| 19 | Comparison of ranked and random allocations .. .. .. .. .. | 89 |

## FIGURES


| | | Page |
|---|---|---|
| 1 | Psychology and economics clusters .. .. .. .. .. .. .. | 44 |
| 2 | Amorphous cluster .. .. .. .. .. .. .. .. .. .. | 45 |
| 3 | Single-linkage hierarchical clustering of psychology cluster .. | 50 |
| 4 | Complete-linkage hierarchical clustering of psychology cluster .. | 51 |
| 5 | Criminology clusters (1950) .. .. .. .. .. .. .. .. | 55 |
| 6 | Criminology clusters (1960) .. .. .. .. .. .. .. .. | 57 |
| 7 | Criminology clusters (1970) .. .. .. .. .. .. .. .. | 63 |
| 8 | Hierarchy from ranked sample .. .. .. .. .. .. .. .. | 76 |
| 9 | Hierarchy from random sample .. .. .. .. .. .. .. .. | 83 |

ABSTRACT

This report describes research carried out as part of the Design of Information Systems in the Social Sciences (DISISS) project. Cluster analysis techniques were applied to a machine readable file of bibliographic data in the form of cited journal titles. The aim was to identify groupings which can be used to structure bibliographic files. Practical applications of the data obtained from clusters of journal title citations were to include the planning of patterns of journal coverage for secondary services, ascertaining information on the structure of the primary literature of a discipline and the provision of data to help validate classification schemes and thesauri. The analyses were carried out on a file of over 60,000 citations taken from social science journals. A special cluster analysis was made of citations taken from criminology source journals at three different time intervals. The report contains a review of clustering techniques applicable to bibliographic data, a detailed discussion of the results of the studies and an evaluation of the practical value of the techniques. In conclusion it was found that the use of cluster analysis is probably not a very practical aid to the design of secondary services in the social sciences. Interpretation of some of the results proved to be complex and often inconclusive, but the techniques nevertheless have important bibliographical applications elsewhere than in the design and evaluation of secondary service coverage.

# 1.0 INTRODUCTION

## 1.1 Statement of aims

The DISISS project consisted of a closely related set of research studies aimed at providing information which could be used to improve secondary services in the social sciences. The aim of this part of the project was to see how far it is possible to use an objective criteria in defining subject areas to be covered by secondary services, as compared with the existing traditional and rather arbitrary divisions.

Since few subject boundaries in the social sciences are clear cut, and because the field is changing all the time, any secondary service which attempts to cover less than the entire field is faced with the problem of defining its scope. Traditional methods of classification have two weaknesses, their inflexibility and their reliance on human judgement. In particular, subject areas involving interdisciplinary work are likely to be fragmented in any sort of classification scheme which is based on a subject hierarchy. This study was an attempt to devise a flexible method of dividing up the social sciences objectively, based on the behaviour of people working in the field. Ideally, we would have liked to separate the literature into clusters such that all items of relevance to any individual would be in a single cluster, but since this is impossible, the aim was to divide the literature into clusters which come as close to this goal as possible.

The main objective of applying cluster analysis to bibliographic data, in the form of journal titles, authors or papers, is to identify groupings which can be used to structure bibliographic files. DISISS deals with data on cited journal titles. Ideally, the data obtained from clusters of journal title citations might be applied in the following ways :

(i) It could help in planning patterns of journal coverage for secondary services. Clustering identifies groups of journals by classifying them into related subject groups. Since the citations used in clustering are user-generated, the clusters provide the designer of a secondary service with a measure by which he can attempt to match the supply of information to the use of information. This applies to existing services as well as providing a basis for new services.

(ii) Journal title clusters provide data on the structure of the primary literature in a discipline. From this general picture:

(a) new fields can be identified, either by comparison with earlier results or because unexpected patterns are displayed;

(b) the need for new or modified services can be seen and action taken if needed;

(c) hypotheses can be developed which may lead to new lines of enquiry and the development of new services.

(iii) Cluster patterns can give evidence on the validity of classification schemes, and on thesauri, which are generally based on an underlying classification.

(iv) Cluster data is relevant not only to the operation and design of secondary services; the data can indicate where the primary literature could be rationalized, or even expanded.

(v) Clustering allows various descriptive studies of the social science literature to be carried out within a firm framework of subject groups. For example, clusters constructed at different points in time can yield information on the growth and obsolescence of subject groups. Similar studies can be carried out on language and country of origin of the journals in each subject group. These studies may be carried out as comparisons between subject groups and within each subject group.

Taken together these five suggested aims constitute a mammoth programme of work. For DISISS it was decided to concentrate on the first and second. The decision was made to cover the whole field of the social sciences, which automatically led to very large data files being required.

## 1.2 Reasons for using citation data

A very large number of citations - over 60000 in all - was collected for the DISISS project, in order to carry out analyses that would, among other things, help to reveal relationships between subjects, languages and countries of publication, to show data distribution, to indicate significant serials, and, most relevant to this report, to define empirical subject boundaries by means of clustering. Citations are by no means perfect indicators of value

or use, but they have a strong positive relationship with both, and have the enormous advantage of being easy to collect. A fuller discussion of the use of citations and citation analyses is contained in DISISS Research Report A3.

DISISS was unusual in gathering references not merely from serials but from monographs also, and in analysing citations to other forms of literature then serials. However, for the purposes of clustering only references made by journals to journals were used, because the scatter of references and citations in and to other forms is so wide that vast amounts of data would have to be collected if meaningful results were to be obtained. This limitation, and the intrinsic limitations of citations as indicators, do not seriously affect the validity of clustering as used in this study.

## 2.0 THE USE OF CLUSTER TECHNIQUES

### 2.1 Survey of previous work

Clustering, grouping, and clumping techniques have been applied to three types of bibliographical data :

(a) citations;

(b) index terms, which may be derived from titles, abstracts or (very rarely) from a text, or may be assigned to a document;

(c) terms in user requests.

Clustering work involving index terms usually has the objective of automatic classification of documents. This is also the objective of some work on clustering documents by citation patterns, but clusters of journal titles are of most interest in dividing the journal literature into groups according to the judgements of scientists themselves as indicated by their citation practices. Work on clustering user terms usually aims to match queries with groups of documents identified in retrieval, which may themselves have been identified by grouping of terms in documents.

Much more work has been done on automatic classification of documents according to index terms than on grouping documents by citation patterns. Important reviews are those by Stevens (1965), Salton (1965), and Stevens, Giuliano and Heilprin (1965). These reviews mention clustering by citations, but deal mainly with grouping of index terms. Important work in grouping index terms is by Salton (1969), Borko (1965), Borko and Bernick (1963), (1964), Dale and Dale (1965), Doyle (1962), (1964), Gotlieb and Kunar (1968), Sparck Jones (1971), and Stiles (1961). Three studies using index terms have been reported by Williams (1966), Wolff-Terroine, Rumbert and Ronatt (1968) and Augustson and Minker (1970). Worona (1969) uses the terms in user requests as well as documents, and develops a query clustering procedure. Interest in classification and grouping of user requests is more recent, and neither this work nor the work on the grouping of documents by index terms will be discussed further in this report.

There is relatively little work to discuss about the application of clustering techniques to citation data. Little attention has been given to the statistical techniques, the requirements which citation data makes, the conditions which apply to the statistical techniques, and the extent to which citation data can meet them. Failure to meet basic requirements gives rise to problems in the interpretation of the results.

Interest in the use of clustering techniques for the organization of bibliographical material has largely arisen in the last ten years. Five studies are of particular interest.

Xhignesse and Osgood (1967) applied an interpoint distance procedure to a citation data matrix consisting of psychology journals; the method produced clusters of interacting journals. The method was tried out on a symmetrical universe of source journals (the 21 source journals also being the only cited journals allowed) and the results obtained appear to be useful.

Price and Schiminovich (1968) used a clustering procedure with a bibliographic coupling measure on a collection of 240 theoretical high energy physics papers. This study was intended as a first step towards producing an entirely automatic classification scheme. Later work by Schiminovich (1971) was designed to overcome some of the failings of the simple approach used in 1968, which appears to have been discarded. They developed a "bibliographic pattern discovery algorithm" using more information on the bibliographic links between documents than is contained in the simple bibliographic coupling measure. This algorithm forms the basis of a method for generating groups of papers and a classification system for documents in the subject area. When applied to a collection of about 30,000 physics documents the groups of documents generated corresponded to recognizable topics even when spread over several conventional classification categories. The generated classification system compared favourably with one used by a journal for indexing its articles.

Carpenter and Narin (1972) have used a cluster analysis procedure in which 288 highly cited journals in physics, chemistry, and molecular biology were grouped into clusters of related journals. The clusters could be identified by national, subject and subdisciplinary divisions.

Narin, Berlt and Carpenter (1972) also investigated hierarchies of journals, using the two journals most cited by each journal in the study, to divide chemistry journals into groups. This first stage might be done by hand. It results in groups containing a manageable number of journals for further clustering. In other words, journals are divided roughly into subject, discipline, or interest groups and cluster analysis can then be undertaken on each group. This is much easier than performing the clustering initially on a large heterogeneous group of journals, partly because of the computational problem of dealing with more than 100 source journal titles in a cluster, and partly because of the difficulty of interpreting clusters

derived from a large, heterogeneous set of titles, covering many subjects. This approach could not be applied to DISISS, because one of the objectives was to investigate the value of cluster analysis in dividing the subject matter of the social sciences, where there is less agreement about discipline boundaries than in the physical and biological sciences.

## 2.2 Clustering terms

A clustering approach or technique is a method of grouping data points without using a preset classification. A clustering algorithm is a description of such a method which is sufficiently well-defined to be programmed or executed manually without further definition. A hierarchical clustering or classification is one in which at the lowest level each data point is a separate cluster and at higher levels each cluster is formed by merging complete clusters from a lower level. At the highest level the whole data set forms a single cluster. Such a clustering can be shown in diagrammatic tree form as a dendrogram.

A multivariate data set is one in which each point is defined by values on a prescribed set of variables. A similarity matrix, S, for a data set is a matrix of positive elements, $S_{ij}$ being the similarity between point i and point j. Usually S is assumed to be symmetric ($S_{ij} = S_{ji}$) with similarities in the range [0,1]. In this case $S_{ij} = 1$ implies that i and j are identical within the terms of the data and the matrix has only $n(n-1)/2$ significant entries since clearly $S_{ii} = 1$. A dissimilarity matrix is the opposite. $S_{ii} = 0$ and the range of the dissimilarities is seldom restricted.

## 2.3 Clustering techniques

Cluster analysis differs from other statistical techniques such as regression or the analysis of variance in that it is not a well-defined technique. The term is applied to a number of widely different approaches, whose only common feature is the objective of identifying groups of points within a given set of data points. This is as far as similarity goes. Some approaches try to detect a single isolated group at a time, which is distinct from the rest of the set. Others try to divide the set into groups to optimize some overall criterion. Yet other methods impose a hierarchical structure on the data. The definition of a cluster, the criterion for optimization or the rules for forming a hierarchy can be varied to stress different qualities required of the clusters: isolation, connectedness, compactness. The purpose for which the clusters are to be used, and hence the requisite qualities and

structure of the clusters, are clearly important when selecting a clustering technique.

The form of the data required by the various methods can be either a multivariate array, where each point is described by its values for a fixed set of variables, or a similarity (or dissimilarity) matrix, where a similarity is specified between each pair of points. It is always possible to convert a multivariate array into a similarity matrix by defining a suitable distance function (e.g. Euclidean) of the variables for a pair of points, but this process has disadvantages. The most important disadvantage is that where the number of variables is small relative to the number of points the similarity matrix is much larger than the equivalent multivariate array. In addition, if it is conceptually reasonable to picture the points in n-dimensional space (particularly if the variables are comparable and reasonably independent) some useful information may be lost in the conversion. In particular the facility is lost for directly identifying a cluster by a hypothetical point (e.g. the centroid of the points in that cluster), which can be a computational or a conceptual aid. The advantages of the similarity matrix appear when there is no convenient fixed set of variables (e.g. when using co-occurences of keywords in their titles to measure the relationship of two articles) or when the variables are not comparable (e.g. height, weight and age). In the latter case the distance function, which will be required for a technique using the multivariate array, may be so complex and time-consuming to evaluate that it is more efficient to determine the similarity matrix once and for all rather than continually to evaluate distances within the main algorithm. Clearly, the form and amount of the data which is available, or which can be collected conveniently, must be also considered when selecting a clustering technique.

Another reason for taking the amount of data to be clustered into account is that clustering techniques vary considerably in the time taken to cluster a given number of points. Any approach which requires complex calculation for its basic step is unsuitable for large amounts of data. For instance, an approach using a complex overall criterion for optimizing the division of a set into clusters will not be efficient if it has to be wholly or largely recalculated whenever a decision has to be made whether to move a point from one cluster to another. In particular, if the multivariate array or similarity matrix is too large to fit into the computer store and some form of paging is required, it is essential to use a technique which does not require the whole matrix for every basic step.

2.3.1 The SCICON algorithm

The clustering method which was chosen is based on a non-hierarchical method developed at SCICON by E.M.L. Beale and M.G. Kendall. This method, referred to as the SCICON method where confusion might arise, assumes that the data is in the form of observations, or data points (in our case cited journals) consisting of measurements on each of a fixed set of variables (in our case citing, or source, journals). If the number of variables is n, these observations are considered as points in an n-dimensional Euclidean space.* At any stage, a point is allocated to one and only one cluster. The criterion used to decide which of two different divisions of the points into a number of clusters is taken, is the root mean square deviation from the centres of gravity of the clusters (see Appendix I ). This is a measure of the average distance of points from the centres of gravity (assuming equal weights for all observations) of the clusters to which they have been allocated. The better allocation will have a lower value for the criterion.

The basic idea is to divide the points into clusters which are so chosen that, for a given number of clusters, the average distance from the points to the centre of the cluster to which they are allocated is minimized. If this criterion is evaluated for divisions into different numbers of clusters, the optimal division will be better for the larger number of clusters. The criterion is therefore not applicable to comparisons of allocations to different numbers of clusters. The procedure used is that, given an allocation to m clusters, each point is examined in turn to determine whether moving it to any other cluster will improve the criterion. If so, it is moved and the next point (in order of submission to the program) is then considered for reallocation. When a full pass through the points performs no reallocations the current allocation is output as the best division into m clusters that

---

* In a Euclidean space, if the ith point $X_i$ is represented as $(x_{i1}, x_{i2}, \ldots, x_{in})$ where $x_{ik}$ is the measurement of observation i on the kth variable, the distance between $X_i$ and $X_j$ is defined as

$$\left( \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

i.e. Pythagoras's theorem extended to n dimensions.

has been found. This may not be the optimal allocation of the points to m clusters that could be found by total enumeration, but this would be impossibly time-consuming for more than a trivial amount of data. The factors which restrict the optimality are the initial allocation, the order in which the data is submitted and the consideration of only one point at a time for reallocation rather than allowing groups of points to be reallocated simultaneously. With the large number of source and cited journals that we have used it is not feasible to alter the last factor because of limitations of computer storage and run time.

The method is non-hierarchical but, in any run of the program, the algorithm produces a sequence of divisions into a decreasing number of clusters. The initial allocation for the m-1 cluster stage is obtained by merging the two closest clusters in the best allocation obtained for m clusters. This procedure means that computation for later stages is often small and that, although not strictly hierarchical, the sequence of cluster allocations obtained often resembles a hierarchical structure.

The allocation of points to clusters at the beginning of the first stage is made as follows. A sequence of cluster centres is formed by first selecting the data point furthest from the centre of gravity of the whole data set as a cluster centre, and subsequently choosing as the next cluster centre the point which maximizes the minimum distance from the point to any of the previous cluster centres. Each point is then allocated to its nearest cluster centre. This procedure has been found by experience to yield quite a good initial allocation and is simple to program and quick to execute. It is advisable to use a higher value for the maximum number of clusters to be considered than is strictly required, to allow a running-in period of several stages. This running-in period will usually counteract any effect of the order in which the data is submitted and the technique of choosing the first set of cluster centres.

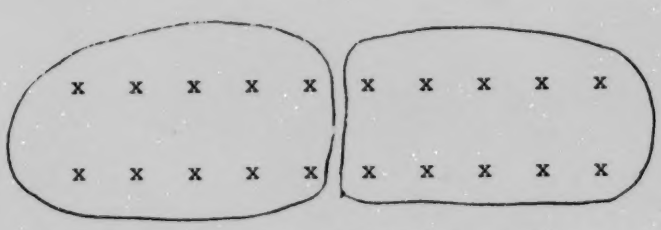
It will be seen that this method is essentially a pragmatic method which experience has shown gives consistently good results with a variety of real life data. It does not require data to follow any particular statistical law,

and freak sets of data can easily be imagined which it would not handle. For instance, in two dimensions, the following data

x x x x x x x x x x

x x x x x x x x x x

appears to fall into two groups, but the algorithm might cluster it as follows.




2.3.2 Reasons for selection of the SCICON algorithm

In DISISS we wished to cluster a large number of social science journals, certainly several hundred. A hierarchical structure, while an obvious advantage in designing automatic retrieval systems for documents, would be an unnecessary restriction in the design of secondary services. However, if a method which does not impose such a structure were to reveal one, this would be valuable information. Social scientists use and cite much material outside their immediate discipline. It seems therefore more reasonable to attempt to divide the set of journals into clusters which optimize some overall criterion than to identify single groups which are distinct from the rest of the set. The latter approach requires a definition of a cluster which is liable to be either so stringent that only a few small groups of journals are sufficiently isolated to satisfy the definition, or so lax that the number of clusters produced by the algorithm is so large that the analysis required after clustering would be enormous.

The data was limited by what could be collected conveniently and at a reasonable cost. Although some data was available from Science Citation Index* the social science source journals covered by this service are mainly in the field of psychology and, to cover the rest of the social science disciplines, it was essential to collect most of the data independently.

---

* Social Science Citation Index was first published too late to be useful to this study.

It was impossible to decide which journals were to be clustered and to collect citations to all of them, from all of them, for three main reasons.

(i) Part of the DISISS project is to identify the important social science journals and since the citations collected are partly for use in this identification, a circular situation arises.

(ii) Copies of foreign and specialist journals are often difficult to track down and it is of interest to cluster these.

(iii) The cost of collecting sufficient citations from all the journals to be clustered is prohibitive.

The alternative procedure is to select a few source journals and collect a large number of citations from each, and then cluster all the journals cited by them (in practice it is reasonable to consider only those journals cited more than a certain number of times). The choice of source journals is dealt with in Section 3. If we regard each source journal as a variable, we have a multivariate array. Each cited journal is a point defined by the number of citations from each source journal. Obviously the number of source journals will be small compared with the number of cited journals, and the variables, all being numbers of citations from a source journal, are comparable.

From these considerations it is reasonable to choose a method which uses a multivariate array to divide a set of points into a number of clusters. The SCICON method satisfies these requirements. It has several other advantages, one being its availability in the form of an efficient computer program. Although, in common with most algorithms which use an overall clustering criterion, it does not guarantee an optimal solution, there is a facility for clustering over a range of decreasing numbers of clusters, and experience with a variety of real data has shown that although the clustering obtained for the first two or three numbers in the range may be definitely sub-optimal, after this initial run-in near-optimal clustering is usually obtained.

Because the change in the criterion (mean square deviation of points from the centre of gravity of their cluster) when a single point is transferred from one cluster to another is simple to calculate using only the position of the point itself and the two relevant centres of gravity (see the Appendix I), the algorithm can cope efficiently with large numbers of points.

The ability to cluster over a range of number of clusters makes it possible to detect whether a distinct hierarchy is present by checking whether a reduction in the number of clusters results in the amalgamation of whole clusters, or in the redistribution between several clusters of points which had been in one cluster. Clusters which remain unchanged over several reductions in the number of clusters are clearly fairly distinct and isolated. Useful subsidiary information which is available for each clustering is the ratio of the distance of each point from its second nearest cluster centre to the distance of each point from its own cluster centre. This is a measure of the adhesion of the point to its cluster. If the ratio is nearly one, the point would fit nearly as well into the second cluster. This ratio might be incorporated in an extended method to allow overlapping clusters. Other useful information is the location, and hence relative location, of the cluster centres, and the mean distance of points in a cluster from its centre, which is a measure of the density or compactness of a cluster.

## 2.4 Comparison with other approaches and algorithms

To compare the SCICON method in detail with all clustering approaches that have been suggested would be an enormous task. This section covers some of the most widely known and some which have been applied to the specific task of grouping journals. It also attempts to highlight some of the main differences in approach and the reasons for these differences.

Over the last thirty years considerable effort has been exerted in developing automatic techniques for dividing a data set into natural or homogeneous groups without previously specifying the groups in any way. This effort has been spread over a number of disciplines: biology (where it is usually known as numerical taxonomy); psychology (e.g. for grouping subjects according to experimental results); linguistics (e.g. for classification of phonemes); information retrieval (e.g. for classification of documents in a collection); and marketing (e.g. to identify groups of competitive products and gaps in the market which might be filled with a new product). The different contexts impose different constraints on the form of clustering required and the data available. Both these factors

impose requirements on clustering techniques and the algorithms which implement them. No single approach could satisfy the full range of requirements made in all circumstances. The following questions highlight some of the differences:

(i) Is a hierarchical structure necessary (e.g. for definition of a tree structure, for efficient automatic information retrieval, or for biological taxonomy) ?

(ii) Is the clustering required to group all the data ? Can outliers be ignored ? Can overlapping clusters be allowed ? Is it required merely to identify individual clusters which are distinct from the rest of the data ?

(iii) In what form is the data ?

  (a) Similarities between data points (e.g. co-occurrence of key-words, numbers of confusions of two phonemes).

  (b) Measurements on comparable variables (e.g. counts of citations from journals to each other).

  (c) Measurements on a set of variables which are not directly comparable (e.g. height, weight and age of individuals).

(iv) How much data is there (20, 100 or 1000 points) ?

The fourth of these questions probably implies a further one. Is the exact clustering of individual points important, or merely the statistical properties of the overall grouping ? An individual plant in a small plant taxonomy is different from the position of an article in a large document collection, where the statistical retrieval performance is the important factor.

The SCICON method is unsuitable if a strictly hierarchical structure is required. It is not obvious whether a hierarchical method is appropriate for classifying social science journals. To impose a hierarchical structure

on data which has no inherent hierarchy is undesirable, but if there are indications of such a structure it is certainly of value to compare the results of hierarchical methods with those of the SCICON algorithm. The most common family of hierarchical methods can be described as follows.

(i) Input the $\frac{n(n-1)}{2}$ similarities between the n points to be clustered.

(ii) Consider each point as a separate cluster.

(iii) Choose the two closest clusters p and q and merge them into a new cluster k which replaces p and q.

(iv) Compute the distance of cluster k from any cluster s, as a function of the distance of p and q from s.

(v) Return to step (iii).

The process stops when the last two clusters are combined. Associated with each merger is the distance between the two clusters when they are merged. The distinguishing factor between members of this family is the distance function computed in (iv). Cunningham and Ogilvie (1972) list 7 variants and Anderberg (1971) discusses 5 of these and adds another. They are listed in the table below.

Table 1

Distance functions suggested for hierarchical clustering

| | |
|---|---|
| 1 | Single linkage (nearest neighbour) |
| 2 | Complete linkage (furthest neighbour) |
| 3 | Average linkage between merged groups (group average) |
| 4 | Average linkage within new group |
| 5 | Centroid |
| 6 | Median |
| 7 | Ward's error sum of squares |
| 8 | Simple average |

Variant (7) in the table is closely related to the SCICON criterion and Anderberg (1971) has experience which suggests that its results when evaluated using the SCICON criterion are good. Compared with the SCICON algorithm its disadvantage is that it requires to start with n clusters where n is the number of data points, and is therefore very time consuming for large data sets. Any alteration to the implementation to counteract this would leave something very similar to the SCICON algorithm, with the disadvantage of not allowing redistribution between clusters to improve the criterion value.

Cunningham and Ogilvie (1972) suggest that on several sets of artificially structured data (3) was consistently good at revealing the structure while (2) and (8) produce similar results (Anderberg (1971) suggests that (4) performs similarly). They found that (5) and (6) tended to distort input data that was strictly ultrametric and thus hierarchical* but performed reasonably on data describing not very distinct clusters in the Euclidean plane.

The function related to the SCICON criterion, (7), performed very well on all types of data except one. The criterion favours clusters which do not differ greatly in size and (7) did not reveal the structure of hierarchical data which represented such clusters. This tendency is not a disadvantage when the aim is to use clusters in the design of secondary services. One large cluster together with a few very small ones would be difficult to interpret and use rationally. However, it implies that the criterion is not suitable for use when such a structure is suspected and the hierarchical position of the individual items is important.

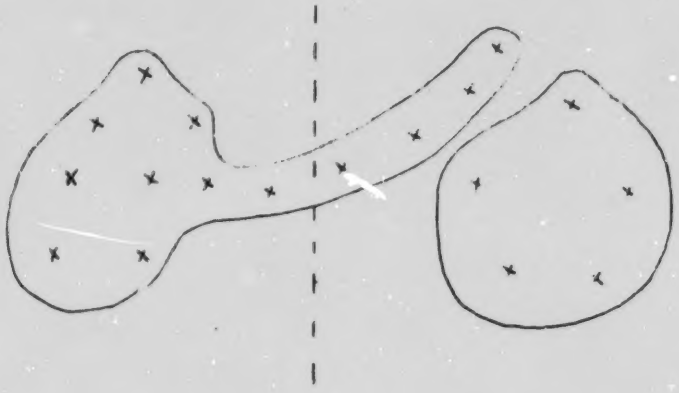
Single linkage (1) is probably the most widely known and used approach, largely because it is very simple to apply efficiently to small or large sets of data. Its disadvantage, which is often considerable when clusters are not

---

* Johnson (1967) develops a correspondence between a hierarchy and similarity matrix satisfying the ultrametric inequality

$$d(x,y) < \max[d(x,z), d(z,y)]$$

very distinct is a tendency towards chaining. Chaining is best illustrated diagramatically.



Although the intuitive way of splitting this data into two clusters might be to divide it along the dotted line, single linkage will tend to form the left-hand cluster and then follow the central chain of points incorporating one at a time, thus producing the two circled clusters. Single linkage can therefore easily produce long straggly clusters in which extreme points are very dissimilar. For some, but not all, applications this is a disadvantage.

Methods (1) and (2) have the conceptual advantage that the resulting clusters strictly maximize intuitive properties of connectedness and compactness respectively. At a given stage of single-linkage clustering, identified by a distance r, all points that can be joined by a chain of points less than r apart are in the same cluster. Any stage of complete linkage clustering the maximum distance between any two points in the same cluster is s. As the algorithms work up the hierarchy, r and s increase steadily. Complete linkage (2) therefore concentrates on compactness and is probably more suited to applications where the overall picture is more important than the positions of each individual. This is borne out by the results in later sections. These two methods in one sense represent extremes between which results from the other five may be expected to fall.

One feature of single and complete linkage which is often stressed is their invariance under all monotone transformations of the similarity matrix. A monotone transformation is one which preserves the ranked order of the elements of the matrix. Johnson (1967) among others has suggested that psychologists only have sufficient confidence in their similarity data to

say that if d(a,b) > d(p,q) then a and b are more similar than p and q. They may not be prepared to concede any significance to the size of the difference between the similarities. However, for large sets of data with similarities restricted to a finite range, as is often the case, this lack of confidence loses its significance. Hubert (1972) suggested that in fact psychologists would be prepared to concede more confidence in their data, to the extent of the rank order of the first differences of their similarities. Transformations which preserve this more constrained rank order are called hypermonotone. He suggests a simple extension from single and complete linkage which is invariant under hypermonotone transformations, though not under a general monotone transformation, defining the distance between two clusters as the sum of the minimum distance between a point in one cluster and a point in the other and the maximum distance between such pairs of points. His experience shows as expected that this produces results intermediate between single and complete linkage. Unfortunately it is more difficult to program efficiently than either.

One point worth making at this stage is the distinction between a method and its algorithmic implementation. Although the eight methods listed in Table 1 can be described as variants of a simple process, often the most efficient algorithm for obtaining the results is by a completely different sequence of steps. Efficiency is affected by two restrictions, time and space, and which is the most efficient algorithm (measured by total cost) may depend on the size of the data and on the particular computer*.

R.F. Ling (1972) has suggested a method which he regards as a generalization of the single-linkage method. As with single and complete linkage his definition of a cluster has a conceptual basis though it is somewhat more complex. In essence, he defines a (k,r) cluster S as follows.

(i) Each point in S is a distance less than r from at least k other points in S [(k,r) bonded].

(ii) Any two points in S can be connected by a chain of points in S in which each link distance is less than r [(r) connected].

If k is chosen as 1, this method yields the single-linkage method. Complete linkage requires clusters of k elements to be (k-1,r) bonded for some r.

---

* Sibson (1973) has published an efficient single-linkage algorithm for very large data sets.

For any fixed k, the set of (k,r) clusters in a set of data forms a hierarchy. However, the hierarchy differs from those discussed earlier in that a merger may be of several clusters rather than just two. To take full advantage of the promising features of this method it would be necessary to compare the hierarchies obtained for different values of k. This would of course be time-consuming, especially as run times increase exponentially with k. A program has been obtained from Ling and it would be interesting to investigate its results on a small data set.

Before leaving the topic of hierarchical methods it is worth pointing out that some of the least desirable features for hierarchical methods listed by Jardine and Sibson (1968) apply only to applications where the exact position of individuals is the important feature. Since the single-linkage method is the only one found to possess these features this is reassuring.

Non-hierarchical methods are considerably less easy to classify and compare than the hierarchical ones. They have often been designed for a particular type of data. Several methods have been designed principally for binary (dichotomous) data of presence or absence variables, and though theoretically applicable to transformed multi-state or continuous data the results may be harder to interpret in the light of the required transformations.

Harrison (1967) describes a program developed in ICI to detect clusters of biologically active compounds using data on chemical structure. It uses binary data and probabilistic measure of cluster significance to detect patches of greater than random density of active compounds. It does not attempt to partition the whole data set into clusters and although an adaptation of the ideas might possibly be appropriate to DISISS the task of programming and determining run parameters by experimentation would have been considerably greater than for the SCICON, method which was already available as a program compatible with the DISISS data.

The broad concept behind most non-hierarchical methods is to start with an initial partition of the data and then move points between clusters to obtain a better partition. Methods differ as to what constitutes a better partition and what techniques are used for improvement. Where an explicit criterion, such as the SCICON criterion*, is used, a hill-climbing algorithm is usually employed, possibly with modifications. Such algorithms allow only

---

* This criterion is sometimes referred to as trace W, the trace of the pooled within groups scatter matrix, as by Friedman and Rubin (1967).

uphill steps, steps which positively improve the criterion. Hill-climbing algorithms are always liable to find local rather than global optima, since it might be better in the long term to go down into a valley to reach a higher peak. The SCICON program uses a hill-climbing algorithm at each cluster level and can only guarantee a local optimum at any level. However, experience has shown that near-optimal results are usually obtained by using a run-in period of about 5 cluster levels. Although absolute optimality could not be checked for the results obtained, consideration of the criterion values after various run-in periods using the pilot study data tends to corroborate this claim. Other explicit criteria for multivariate data have been discussed by Anderberg (1971), Friedman and Rubin (1967), and Scott and Symons (1971). Most of these have the disadvantage of requiring complex and time-consuming calculations in testing whether moving a point will improve the criterion or not. Very little published work has been done using them. Rubin (1967) introduced another criterion using a concept of average object stability, where the stability of an object in a cluster is a measure of the similarities of a point to the other points in its cluster compared with its similarities to points in another cluster. This criterion is based on a similarity matrix rather than directly on multivariate data, and is comparable for division into different numbers of clusters. This last feature means that Rubin's modified hill-climbing algorithm finds an optimum number of clusters as well as an optimum partition into a particular number of clusters. However, his criterion also involves a variable parameter, and runs for several values would be necessary for useful results*. He suggests a number of tactics to apply in an attempt to improve on local optima but remarks that they are only occasionally helpful. This method, which would be very complex to program, has been little studied. This neglect may be due to its complexity rather than to any obvious defect.

Several people have suggested algorithms which while not optimizing an explicit criterion implicitly use one very similar to the SCICON criterion. Anderberg (1971) discusses several of those which, given the centroids of a partition, allocate points to their nearest centroid, recompute the centroids and recycle. The algorithmic details vary, and different methods of improving on local optima or optimizing the number of clusters are suggested, but none of them shows any obvious advantage over the SCICON algorithm.

---

* A paper by Gitman (1972) suggests a similar approach involving no run-time parameters.

There are a few published examples of attempts to group journals. These have tended to use rather ad hoc methods which might be difficult to apply to large amounts of data. Brief descriptions of three attempts follow.

Xhignesse and Osgood (1967) used citations within a group of 21 psychology journals to obtain a similarity matrix (the exact calculation used is not described). A method developed by Shepard (1962) was applied to this matrix to represent the journals in multidimensional Euclidean space in such a way that the rank order of the similarity elements was preserved. This is not a clustering technique; it is a method for obtaining a Euclidean multivariate representation of data which naturally produces a similarity matrix. After applying this process Xhignesse and Osgood used an arbitrary distance to define overlapping clusters of journals. The exact procedure is not stated but it is probably based on defining clusters as groups of journals lying within spheres of a certain diameter. It is most unlikely that this method which is rather arbitrary and ill-defined (at least in the published paper) would be applicable to large amounts of data.

Parker, Paisley and Garrett (1967) found clusters among 68 journals in the field of communications. Unfortunately they did not describe the clustering method in detail and omitted the relevant references from their bibliography. Citations were taken from 17 source journals and co-occurrence of citations to pairs of journals from the articles in these sources was used as a measure of similarity between the journals. This measure is almost an inverse of bibliographic coupling as introduced by Kessler (1963a) and discussed later in this section. A disadvantage of this measure is that, without any adjustment for the level of citation to each journal, pairs of highly cited journals are almost certain to have high similarity. The clustering approach appears to identify a tight cluster of highly related journals, and then discard the journals not in that cluster from consideration. It is difficult to see what application of these clusters is possible especially if journals from more than one field are considered.

Carpenter and Narin (1972) have followed their earlier study (Narin, Carpenter and Berlt, 1972) of the structure of the journal populations in scientific disciplines through relative citation levels between journals, by clustering journals according to citation data. Their clustering approach involves a similarity matrix and a simple hill-climbing method to optimize

an overall criterion. They do not specify whether the choice of the number of clusters is internal to the algorithm or chosen beforehand. They have been unable to decide on a best similarity measure or clustering criterion and, making the reasonable assumption that clusters which are detected by several different methods are definitely clusters, they combine the results using several similarity measures and clustering criteria. This is done by producing a new similarity matrix in which the number of times journal x appears in the same cluster as journal y is the similarity between the two journals. A single-linkage clustering is performed on this matrix to obtain the final clusters. This combined procedure is an admirable approach; although it would be difficult to implement on the whole of the DISISS data, it has many advantages for smaller samples.

The basic data used by Carpenter and Narin is cross-citation between all the journals under consideration, making use of Science Citation Index tapes. It was not possible to reproduce their analysis using social science data because the Social Sciences Citation Index was not available at the time this experiment was carried out. It might be of interest to use their methods of analysis for the source journals in the ranked and random lists for the main DISISS citation study. However, it is possible that 95 journals spread over the whole of the social sciences may be too small a set to achieve results that are not self-evident anyway.

Finally a few comments can be made on some attempts to group documents, usually journal articles, using citations made by these documents.

Preparata and Chien (1967) suggest an algorithm for solving the problem of arranging documents (or document descriptions) in sequence to minimize, in some sense, the distance between similar articles. This could be used to minimize access time for documents retrieved in response to a query, where the collection is held on magnetic backing store of a type that requires movement of the reading heads over the file.

They consider a similarity matrix in which the similarity between two documents is 1 if either document cites the other, and 0 otherwise. The algorithm uses a hill-climbing approach to minimize the total distance between all pairs of documents with similarity 1. Although the details are very different the approach is related to a restriction to one dimension of the scaling method of Shepard (1962) used by Xhignesse and Osgood (1967) and

referred to earlier. Despite the title of the report* by Preparata and Chien this is not clustering in the usual sense of the word. There is no other obvious application of the solution to their problem, and in a computer system it would probably be possible to reduce access time further by an arrangement of the collection specific to the particular hardware.

Kessler (1963a) introduced the concept of bibliographic coupling between articles and used it to define two criteria for grouping papers. The bibliographic coupling between two papers is the number of items cited by both papers. High coupling tends to indicate a strong relationship between articles because they refer to common work. As it stands, this measure is not suitable a as a similarity measure for the clustering techniques described in the earlier part of this section; papers which cite heavily are more likely to be highly coupled with other papers, and the coupling of a paper with any other is limited by the number of citations it gives. No suitable modification of the concept to counteract this effect is immediately obvious, but the approach might be interesting to investigate. Kessler uses the concept to generate two types of group. The first type of group is generated from a triggering paper and consists of all papers in the set under consideration which have at least n citations in common with the triggering paper. The second type of group is one in which each paper has at least n citations in common with every other paper in the group. This second condition is a stringent one, somewhat related to that suggested by Van Rijsbergen (1970), and seems to have been largely ignored by Kessler in later papers (1963a, 1965). Using the first criterion any paper can be used as a triggering paper to generate a group of papers from the set under consideration.

Not all papers will necessarily trigger off a group at all and in that case they will not be in any generated group since they cannot share n citations with any other paper in the set. However, the groups which are generated will certainly overlap, without overlapping groups necessarily being identical. This form of grouping or clustering is less suitable for DISISS than for designing a classification system to be used explicitly or implicitly within an information retrieval system.

---

* F.P. Preparata and R.T. Chien (1967) Report R-349 Co-ordinated Science Laboratory, University of Illinois, May 1967. On clustering techniques of citation groups .

Schiminovich (1971) found the simple idea of bibliographic coupling expressed as a numerical value unsatisfactory for developing a fully automatic classification system. He used both the numerical value and the information of which citations contributed to this value to generate a sequence of groups of papers which should converge to form a bibliography which could define a class in a classification. He used Kessler's idea of a triggering paper but realized that the trigger need not actually be a paper with its list of citations. Any list of citations could serve as a trigger, e.g. a user-provided bibliography, or a group of papers produced by an earlier stage of his automatic process. The pattern discovery algorithm which forms the basic step of his method could be adapted to cluster documents using index terms or word content and could probably be usefully applied to completely different problems in different fields.

## 3.0 DATA FILES

### 3.1 Files used

Three separate files of citations were used.

(i) A pilot study.
(ii) Citations from criminology journals.
(iii) The main DISISS citation file.

The pilot study was used to devise techniques for collecting the other data and to check the clustering method.

The DISISS project used citation data for a variety of analyses of the citation practices of social scientists, one of which was to cluster social science journals. For clustering it is desirable to adopt one of two strategies:

(i) collect equal numbers of citations to journals from each source journal;
(ii) collect all citations (or every nth) from all articles (or every mth) over a fixed time period.

For some of the other analyses citations to all forms of material are required, in the proportions in which they occur, and in such a way that numbers of citations from different source journals can be combined. In this case two reasonable strategies would be:

(i) collect equal numbers of citations from each source journal, and count total number of citations from each source journal;
(ii) collect all citations (or every nth) from all articles (or every mth) over a fixed time period.

The second strategy is identical in both cases, whereas the first strategy would involve identifying two different but overlapping sets of citations to journals. The second strategy is also easier to apply in practice since collectors can work independently and without keeping an exact count of citations as they work. For the main DISISS file it was therefore decided to collect every citation from every article in 1970

for source journals from which citations were collected by hand. Citation collection is time-consuming and costly and since magnetic tapes of the citations in Science Citation Index for one quarter of 1971 were available it was decided to make use of this data source where possible. For source journals on these tapes all citations for one quarter of 1971 were available. It was necessary to use a weighting factor when combining citations from the two sources.

The choice of source journals could have considerable effect on any analyses of citation data. Some thought has been given to this problem (see DISISS Working Paper 5) and it was decided to select source journals in three ways.

(i) Select the fifty journals most cited in the pilot study.
(ii) Select a random sample of fifty journals from CLOSSS*.
(iii) Select key journals in a specified subject area. The main DISISS project looked at several subject areas. Only one of these, criminology, was used for the cluster analysis. In fact three separate criminology files were used: citations from 1970 journals, 1960 journals and 1950 journals. These were used to examine trends.

The data elements required for clustering are source identification and cited title for every citation in a journal. Each title must be identified by a unique code. Identification of source jounals is simple, since codes can be given to the source titles. For cited titles the problem is less trivial since a title may be cited in several forms e.g. J. Exp. Psych., J. of Experimental Psychology, J. Exp. Psy. Since each title in CLOSSS was allocated a unique 5-digit number for book-keeping purposes, it was decided to use the CLOSSS number as the identification code for both source and cited journals. Cited titles not in CLOSSS (i.e. non-social science journals) were given a code consisting of 1 letter followed by 4 digits (a dummy CLOSSS number). Allocation of CLOSSS numbers cannot be achieved fully automatically as visual recognition of title variants is necessary to some extent. However the process only involves a small amount of manual checking and coding since only one occurrence of each title variant requires visual checking.

After allocation of CLOSSS numbers to cited titles, a simple counting procedure was required to produce the basic data matrix. This matrix consists of a row for each cited title, in which the elements are the number of citations from each of the source journals.

* CLOSSS = Check List of Social Science Serials. This list has been accumulated by DISISS and contains about 5,000 titles.

### 3.1.1 The pilot study data

During the summer of 1971, citation data was collected from 17 source journals in the main subject areas of the social sciences. This data is discussed in DISISS Working Paper No. 5 from a descriptive point of view and summarized in Table 2. This data was used to test the program for the SCICON algorithm for programming errors and to investigate the performance of the algorithm on this type of data. The effects of various treatments of the data on the clusters produced was also studied. In addition some of the data has been used with three other clustering algorithms :

(i) single-linkage hierarchical clustering,

(ii) complete-linkage hierarchical clustering,

(iii) an algorithm suggested by C.J. Van Rijsbergen (1970).

These algorithms were selected for simplicity of application, and, because they represent a variety of approaches, comparison of the results obtained with those from the SCICON method is of interest.

The pilot data consisted of approximately 5000 citations from 17 source journals. Several of the cited journals were omitted from the clustering runs because they were cited by only one source journal. Such titles must logically be in the same cluster as the relevant source journal and manual addition can be made after the runs. The number of titles used in the clustering runs was 115, with a further 98 being cited by only one source.

### 3.1.2 Citations from criminology journals

The study of criminology literature is described fully in DISISS Research Report A3. One purpose was to see how primary literature changes with time. For this purpose citations were collected from criminology journal sources for 1950, 1960 and 1970. As far as possible the same sources were used for each year. As well as mainstream criminology journals, related journals were scanned for articles relevant to criminology.

Table 2

Source journals for the pilot study

| Source | Total citations | Mean Citations/ article | Self citations | Self citations as percentage of total citations | Number of times cited by other source journals |
|---|---|---|---|---|---|
| American Anthropologist | 504 | 15 | 61 | 12 | 6 |
| American J. of Sociology | 379 | 9 | 18 | 5 | 23 |
| American Political Science Review | 100 | 14 | 11 | 11 | 6 |
| American Psychologist | 128 | 3 | 18 | 14 | 5 |
| American Sociological Review | 37 | 12 | 2 | 5 | 68 |
| Arcives Europeenes de Sociologie | 176 | 22 | - | - | 4 |
| British J. of Politics and History | 342 | 20 | 7 | 2 | - |
| British J. of Criminology | 273 | 9 | 6 | 2 | 2 |
| British J. of Psychology | 490 | 13 | 40 | 8 | 2 |
| British J. of Sociology | 517 | 21 | 14 | 3 | 5 |
| Economica | 291 | 10 | 25 | 8 | 13 |
| Economic Journal | 295 | 13 | 29 | 10 | 33 |
| Parliamentary Affairs | 167 | 4 | 3 | 2 | - |
| Psychologische Forschung | 421 | 52 | 15 | 3 | 3 |
| Revue Economique | 401 | 9 | 19 | 4 | - |
| Social Service Quarterly | 214 | 5 | 4 | 2 | 2 |
| Sociological Review | 182 | 16 | 4 | 2 | 5 |
| TOTALS | 4917 | 247 | 276 | 5.6 | 177 |

To reduce the volume of data, only every fifth citation was recorded. Table 3 lists the sources and the number of citations to other journals recorded. In the table a dash means that citations were not collected from the journals in that year, almost invariably because it was not published. A zero means that the volume was examined but no articles with citations were found relevant to criminology.

3.1.3 Citations from the ranked and random lists of sources

As part of the main DISISS programme 52,889 citations were collected from 141 source journals. The choice of journals and the methods of collection are described in DISISS Research Report B5. For the cluster analysis two lists of source journals were used.

(i) A ranked list. These were the 52 journals found to be most heavily cited in the pilot study (see Section 3.1.1). They are listed in Table 4.

(ii) A random list. These were 47 journals chosen at random from the Check List of Social Science Serials (CLOSSS). They are listed in Table 5.

These two sets of data were used to see how much the choice of source journals affects the results of a study based on citation data. Citation data was obtained from all the journals on the random list and from 48 of the 52 journals on the ranked list. Two journals appeared on both lists. The data for 60 journals consisted of every citation for 1970 gathered by hand. For the remaining 33 journals the data consisted of every citation on the Science Citation Index tapes for the first quarter of 1971. Thus different numbers of citations were gathered from each source and to each cited journal. The treatment to minimize this bias is described in Section 3.2.

Considerable effort was put into collecting data with the minimum quantity of errors, but inevitably the data contains some errors. The main difficulty lies in reconciling different abbreviations of the same cited journal title. The procedure used was a preliminary computer scan,

Table 3

Citations from criminology journals to other journals

| Source | 1950 | 1960 | 1970 |
|---|---|---|---|
| Acta Criminologica | - | - | 5 |
| Adolescence | - | - | 31 |
| Am. J. Correction | - | - | 3 |
| Am. J. Orthopsychiatry | - | 107 | 0 |
| Am. Sociological Rev. | 57 | 83 | 46 |
| Annales Internationales Criminologie | - | - | 41 |
| Brit. J. Addiction | 11 | - | 60 |
| Brit. J. Criminology | 11 | 17 | 21 |
| Brit. J. Psychiatry | - | - | 0 |
| California Law Rev. | - | - | 14 |
| Crime and Delinquency | - | 4 | 15 |
| Criminal Law Bulletin | - | - | 15 |
| Criminal Law Rev. | - | 45 | 35 |
| Criminologist | - | - | 2 |
| Howard J. Penology | 4 | 5 | 6 |
| Int. J. Social Psychiatry | - | - | 30 |
| Issues in Criminology | - | - | 28 |
| J. Criminal Law, Criminology & Police Sci. | 34 | 64 | 21 |
| J. Research in Crime & Delinquency | - | - | 36 |
| J. Social Issues | 23 | 18 | - |
| Kriminalistik | - | - | 35 |
| Medicine, Science & Law | - | - | 63 |
| Medico-Legal J. | 5 | 11 | 4 |
| Policia Espanola | - | - | 33 |
| Prison Service J. | - | - | 1 |
| Probation | - | - | 1 |
| Psychiatry | 35 | 51 | 0 |
| Rev. Int. Criminologie et de Police Technique | 19 | 4 | 18 |
| Security Gazette | - | - | 0 |
| Social Casework | - | 28 | - |
| Social Forces | - | 99 | 53 |
| Social Work | - | - | 24 |
| Univ. Chicago Law Rev. | 19 | - | 11 |
| TOTAL | 218 | 536 | 642 |

followed by hand checking. On occasion it was impossible to judge whether two abbreviations refer to the same title. The SCI data was worse than the hand-collected data. Fortunately ambiguities were rare and it is safe to assume that data errors have not made any significant difference to the results.

Table 4

Ranked list of source journals

Administrative Science Quarterly
Africa
American Anthropologist
American Antiquity
American Economic Review
American Journal of Orthopsychiatry
American Journal of Psychology
American Journal of Sociology
American Psychologist
American Political Science Review
American Sociological Review
* Archiv für Psychologie
Archives de Psychologie
British Journal of Criminology
British Journal of Psychiatry
Canadian Journal of Economics
* Canadian Journal of Political Science
Canadian Journal of Psychology
Child Development
China Digest (now Peking Review)
Economie Appliquée
Economic Journal
Economica
Econometrica
Human Organization
International Journal of Social Psychiatry
Journal of Abnormal Psychology
Journal of Comparative and Physiological Psychology
Journal of Consulting and Clinical Psychology
Journal of Experimental Psychology
* Journal of Genetic Psychology
Journal of Personality
Journal of Political Economy
Journal of Psychology
Journal of Social Psychology
Journal of the Royal Statistical Society, Series A
Journal of Verbal Learning and Verbal Behaviour
Manchester School of Economic and Social Studies
Monthly Labour Review

Psychiatry
Psychological Bulletin
Psychological Review
Psychonomic Science
Public Opinion Quarterly
Quarterly Journal of Economics
* Quarterly Journal of Experimental Psychology
Revue d'Économie Politique
Review of Economic Studies
Review of Economics and Statistics
Social Forces
Zeitschrift für (angewandte) Psychologie
Zeitschrift für experimentelle und angewandte Psychologie

---

* No citations were collected from the four journals marked with asterisks.

Table 5

Random list of source journals

Aconcagua

Acta Psychiatrica et Neurologica Scandinavica

Activist

* Administrative Science Quarterly

Adolescence

Agrarwirtschaft

Algemen Nederlands Tijdschrift voor Wijsbegeerte en Psychologie

American Federationist

American Journal of Correction

American Vocational Journal

Anglo-Norwegian Trade Journal

Anglo-Soviet Journal

Architecture West Midlands

Arrive

British Journal of Sociology

Bulletin de l'Institut Internationale d'Administration Publique

Cahiers Ferdinand de Saussure

Eastern Librarian

Filosofska Misul

Français Moderne

Geographical Analysis

Head Teacher Review

Human Development

International Journal of American Linguistics

Journal of Clinical Psychology

Journal of Experimental Research in Personality

Journal of the Institute of Education of the Universities Durham and Newcastle

Journal of Psychosomatic Research

Liberal Education

Library Trends

Linguistic Inquiry

Music in Education

Neuren Sprachen

Occupational Psychology

Operational Research Quarterly

Problemi Della Sicurezza Socials

* Review of Economic Studies

Revue de Phonètique Appliquée

Rivista di Diritto del Lavoro

Royal Society for the Promotion of Health journal
Scottish journal of political economy
Smokeless air
Southern economic journal
Soviet geography
Taxation
Tijdschrift voor Economische en Sociale Geografie
West African journal of education

---

* Two journals appear in both the ranked and random lists of source journals.

## 3.2 Treatment of data before clustering

For each of the samples, the raw data consisted of observations of the number of citations from each of the source journals to each journal title which is cited. Consider this as a matrix in which each row is a data point, representing a cited journal by the number of citations to it from each of the source journals. The number of columns of this matrix is the number of source journals (variables) and the number of rows is the number of journal titles cited (data points).

### 3.2.1 Reduction in size of the data matrix

(i) Reduction in the number of rows

The time taken by the clustering algorithm increases rapidly with the number of data points. Since the journals which are seldom cited are not important in the context of clustering by citation pattern, the first step is to drop all journals which are cited by only one source journal. For our purposes such a journal must logically belong to the same clusters as the journal which cites it and can be added manually. It is possible to allocate titles manually to their nearest cluster, without affecting the degree of optimality of the clustering seriously.

(ii) Reduction in the number of columns

One approach would be to ignore columns with small entry totals (i.e. source journals which give few citations). A slight risk is run of failing to find small clusters concentrated on the ignored source journals.

Another approach is to analyse the data by principal component analysis prior to clustering. This procedure successively identifies linear combinations of the original columns which account for the maximum variance not explained by the previous linear combinations. Since maximum variance will tend to give maximum cluster differentiation it might in theory be possible to select a somewhat smaller number of these linear combinations (principal components) as variables while still retaining most of the variance in the data as represented by the original variables. The variables will no longer be identified with individual source journals but with particular linear combinations of the citations from these journals.

To test this theory principal component analysis was used on the pilot

data both before and after normalization. The results in both cases were that all the source journals made a substantial contribution to the overall variance and to eliminate even a single principal component would have noticeably altered the clustering results. Consequently principal component analysis was abandoned and not used with the other sets of data.

3.2.2 Normalization, self-citation adjustment and scaling

For a number of reasons the raw data is unsuitable for clustering directly. The dominant effect is that the data points tend to be allocated in the Euclidean space as in the diagram below, which is an attempt to represent a space of many dimensions on a two-dimensional sheet of paper.

The outlying points are the highly cited titles and the group in the centre are the remainder. The clustering algorithm would produce one very large cluster and a number of small ones, often consisting of a single journal*. Although this is reasonable given the data, it is not the sort of cluster pattern which is required. For our purposes a journal which is cited 10 times by A and 5 times by B would reasonably fall into the same cluster as a journal cited 100 times by A and 50 times by B. It would therefore be logical to take the proportion of the citations to a source journal from each of the source journals as the data. This is achieved by dividing each matrix element by the

* When 105 of the pilot study cited titles were analysed into 15 clusters using the raw data, 75 journals were grouped together and 8 of the other clusters consisted of a single journal.

total of the original elements in its row, and can be thought of as a form of row normalization.

Another form of adjustment which should be considered is scaling of the variables. This can have a large effect on the clustering as shown by the diagram below.

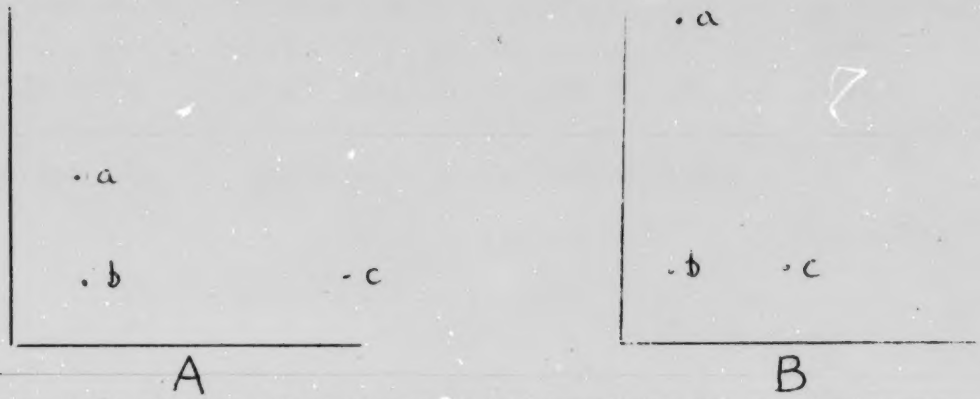



Diagram B is obtained from Diagram A by stretching the vertical axis and compressing the horizontal axis or taking $x' = ax$ $(a < 1)$ and $y' = by$ $(b > 1)$. This scaling of the variable however alters the cluster pattern. In Diagram A, a and b would cluster together at an earlier level than c would cluster with either; but in Diagram B, b and c would cluster before a would cluster with either.

Two simple scaling procedures which could be applied are

(i) scaling by variable totals

(ii) scaling by standard deviation.

The first scales a column down in proportion to the number of citations from the relevant source journal, thereby increasing the importance of source journals which give few citations. In effect it says that if the same number of citations had been collected from each source this is how they would have been distributed, extrapolating directly from the numbers that were collected.

The second procedure gives each column of the scaled matrix the same sample variance. The effect of this on the original correlated variables is

somewhat indeterminate in general. However, it might be a reasonable approach on the uncorrelated variables obtained by principal component analysis when the total variances on individual variables (off-diagonal entries of the covariance matrix being zero).

### 3.2.3 Choice of data adjustments

Apart from the row normalization, which is essential to our purpose, the only way to select which adjustments to make is by experiment. The pilot study data was used for this experimentation. The results of the pilot study are given in the next section. From these results it was decided to use no other adjustments expect for self-citations. The method used for self-citations is described in Section 4.1.5.

## 4.0 THE PILOT STUDY

The pilot study data was used to test the program and to experiment with the various treatments of the data. Clustering runs were made using several treatments of the data:

(i) unmodified,

(ii) with cells divided by row and column totals,

(iii) with cells divided by row totals only,

(iv) with cells divided by row totals and a constant added to each non-zero cell,

(v) as (iii) with self-citation cells reduced by 25 per cent prior to division by row totals,

(vi) as (iii) with self-citation cells which were the highest cell in their row reduced to the next highest cell in the row, this adjustment being made prior to division by row totals.

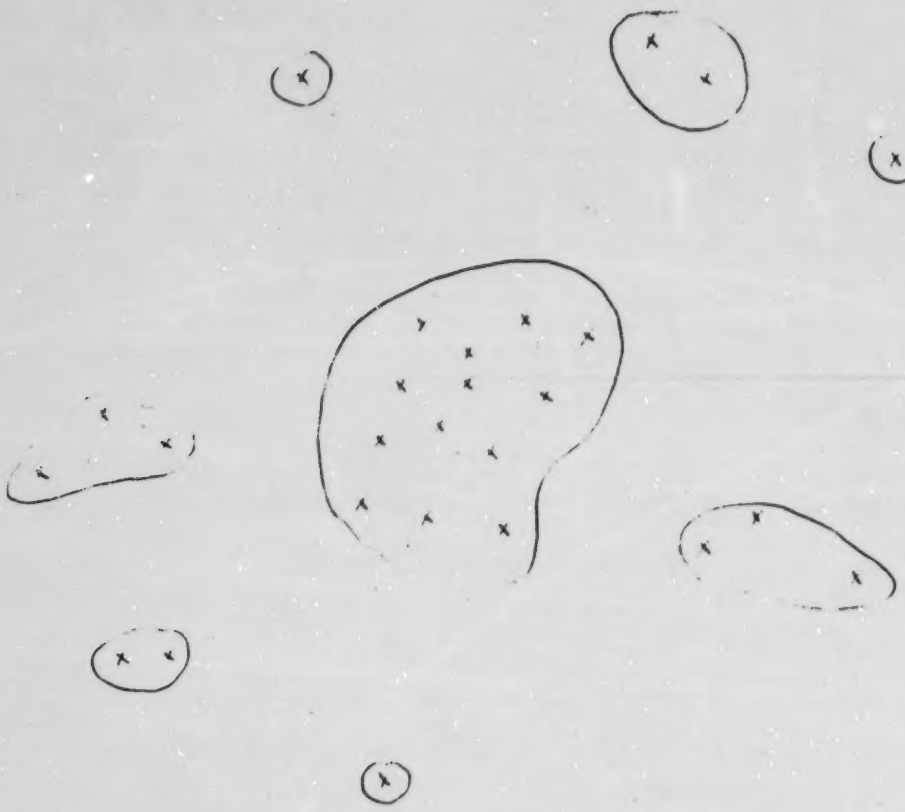
### 4.1 Results

Brief results are given here for most of the data treatments listed above, together with more detailed results for the most promising treatments.

#### 4.1.1 Unmodified data

This run demonstrated the necessity for treating the data in some of the ways suggested in section 3.2. When the journals were grouped into 10 clusters, 80 of the 115 journals appeared in one cluster centred near the origin (the typical journal of this cluster would be cited between .03 and .91 times by each of the sources).

The remainder were the most highly cited journals in the sample, including 11 of the 17 source journals. The data and the clusters produced are of the form shown in the diagram are on the next page.



The outlying points represent the highly cited titles which have high values in one or more variables.

4.1.2 With cells divided by row and column totals

It seems reasonable to adjust for differences in the number of citations collected from each source to approximate the situation where the same number of citations is collected from each source. However, from the run it appeared that this procedure gave too much weight to citations from journals from which fewer citations were collected. Since on the whole these journals were the less important ones, it seemed unreasonable that they should function as discriminators between clusters rather than the more important titles which provided more citations.

In theory it would seem reasonable to adopt one of two strategies when collecting citations for clustering:

(i) collect strictly equal numbers of citations from each source journal;

(ii) collect all citations (or every nth) from all articles (or every mth) for a year, and use the implicit weighting as a

measure of the importance of a source for discriminating between clusters.

The strategies will clearly produce different clusterings but at least underlying properties of the data are known and results can be considered in the light of these features. A possible argument against the second is that journals have widely different numbers of citations per article. In this study the second strategy was attempted, but unfortunately some of the sources had not been in existence long enough, and for some the volume of data to collect was too much to complete in the time allowed. Despite these short comings it appears better not to adjust for the different numbers of citations from the source journals, but to allow the implicit weighting to have an effect on the clusters.

4.1.3 With cells divided by row totals only

More extensive runs performed on this form of the adjusted data than for the others. The 115 journal titles were grouped into numbers of clusters in the range 25 to 3. The full results are given in DISISS Working Paper 11. Here is a summary.

One approach is to look at the individual journals and see which cluster they are in for different numbers of clusters. In Working Paper 11, the journals in each of the clusters at the 3-cluster levels are traced through some of the cluster levels (the levels were chosen arbitrarily). Of these three clusters, one is clearly a psychology cluster of 34 journals and another an economics cluster of 21 journals. The third cluster is less easy to classify and contains 60 journals. Table 6 lists these three clusters. If we consider the first two clusters, it is clear that most journals have one of a small number of traces. Given these trace codes we can group the journals according to their trace codes as in Figure 1. The traces can be used to display a simple network for each of these clusters.

Table 6

Clusters derived from pilot data

| PSYCHOLOGY CLUSTER | |
|---|---|
| | Code |
| Acta Psychologica | A |
| Am.J.Psychology | A |
| Annals Mathematical Statistics | - |
| Annals N.Y.Acad of Sciences | A |
| Arch. für gesamte Psychologie | B |
| Arch. de Psychologie | A |
| Br.J.Educational Psychology | C |
| Br.J.Medical Psychology | C |
| Br.J.Psychiatry | C |
| Br.J.Psychology | A |
| Bull.Br.Psychological Society | A |
| Canadian J.Psychology | A |
| Genetic Psychology Monographs | A |
| Harvard Law Rev. | A |
| J.Abnormal & Social Psychiatry | B |
| J.Comparative Psychology | D |
| J.Comparative & Physiological Psych. | A |
| J.Educational Psychology | D |
| J. Genetic Psychology | A |
| J. Neurology, Neurosurgery & Psychiatry | C |
| J.Neurophysiology | A |
| J.Personality | B |
| J.Psychology | - |
| J.Social Psychology | A |
| J.Operations Research Society of Am. | - |
| Nature | A |
| Occupational Psychology | A |
| Philosophical Studies | B |
| Psychol. Arbeiten | B |
| Psychological Bull. | A |
| Psychological Monographs | A |
| Psychological Rev. | B |
| Psychologische Forschung | B |
| Studium Gen. | B |
| **ECONOMICS CLUSTER** | |
| Am.Economic Review | A |
| Annals Am.Academy of Political & Social Sci. | B |
| Arch. für Sozialwissenchaft | B |
| Australian Q. | B |
| Bull.Oxford Inst. of Statistics | C |
| Canadian J. | - |
| Econometrica | C |
| Economic Journal | C |
| Economic Weekly (Bombay) | - |
| Economica | D |
| Ekonomisk Tidsskrift | A |
| J.Political Economy | C |
| J.Opt.Society of Am. | - |
| J. Royal Statistical Society | C |
| Manchester School | - |
| Monthly Labor Rev. | B |
| National Institute Economic Rev. | - |
| Q.J.Economics | A |
| Rev.Economics & Statistics | D |

Table 6 (Cont.)

| ECONOMICS CLUSTER | |
|---|---|
| | Code |
| Rev.Economic Studies | A |
| Survey of Current Business | B |

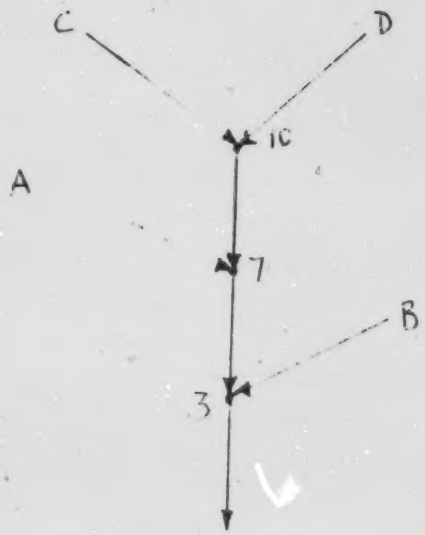
| AMORPHOUS CLUSTER | |
|---|---|
| | Code |
| Actualite Économique et Financière | - |
| Administrative Science Q. | F |
| Africa | D |
| Am.Anthropologist | A |
| Am.Behavioral Scientist | A |
| Am.J.Orthopsychiatry | B |
| Am.J.Psychiatry | C |
| Am.J.Sociology | C |
| Am.Museum of Natural History Anthrop.Papers | B |
| Am.Political Science Rev. | - |
| Am.Psychologist | D |
| Am.Sociological Rev. | C |
| Annals | B |
| Année Sociologique | - |
| Arch.Européenne de Sociologie | G |
| Australian and New Zealand J.Sociology | E |
| Australian Outlook | D |
| Br.J.Criminology | C |
| Br.J.Sociology | A |
| Cambridge Journal | D |
| China Weekly Rev. | B |
| Colliery Guardian | A |
| Comparative Studies in Society & History | E |
| Current Sociology | B |
| Encounter | G |
| Esprit | D |
| Geographical Rev. | B |

| AMORPHOUS CLUSTER | |
|---|---|
| | Code |
| Human Organization | A |
| Human Relations | A |
| Industrial and Labor Relations Rev. | D |
| Int.J.Social Psychiatry | - |
| J.Criminal Law and Criminology (SI) | - |
| J. Experimental Psychology | - |
| J.Philosophy | A |
| J.Social Issues | A |
| J.Am.Statistical Assen. | - |
| Kyklos | G |
| Listener | - |
| New Statesman | D |
| Oxford Economic Papers | - |
| Pacific Sociological Rev. | A |
| Parliamentary Affairs | E |
| Proc.Aristotelian Soc. | A |
| Psychiatry | E |
| Public Opinion Q. | B |
| Review | G |
| Rev.d'Économie Politique | F |
| Rev.Économique | F |
| Rev.Française de Science Politique | F |
| Rev.Int. de Sociologie | B |
| Science | E |
| Social Forces | D |
| Social Problems | - |
| Social Service Q. | E |
| Sociological Q. | B |
| Sociological Rev. | A |
| Sociometry | B |
| Time | - |
| World Politics | G |
| Yale Law Journal | C |

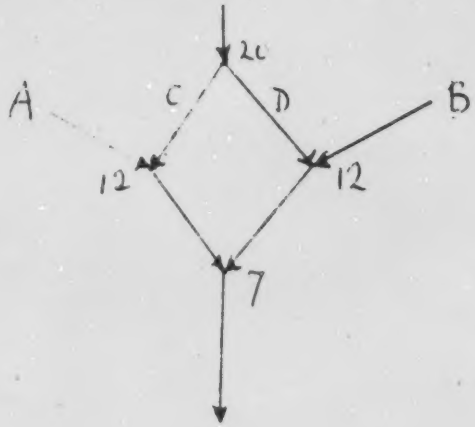
Figure 1

Psychology and Economics Clusters

Psychology Cluster




Economics Cluster

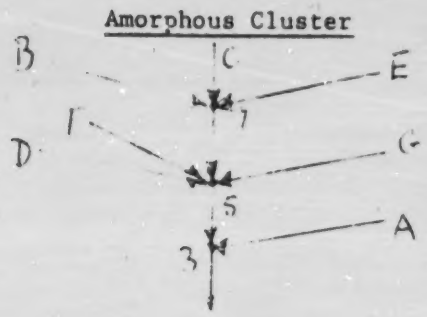



The levels attached to these diagrams have been taken from the traces. More accurate, though not necessarily more helpful, levels could be obtained from the original clustering print-out. Comparison of levels gives a rough guide to the relative cohesion of clusters.

For example, the economics cluster is a separate entity from level 7 downwards and is therefore more cohesive than the psychology cluster, which is in two sub-clusters at level 7 which do not combine till level 3.

Unfortunately the third, rather amorphous, cluster does not lend itself so easily to this type of analysis. By restricting the trace to level 10 and below a good proportion of the journals can be grouped by trace code (see Figure 2), and the corresponding network is shown below. It must be remembered that the structure of this network is less well-defined than the previous two because of the restriction on the trace codes.

Figure 2




Considering these subgroupings subjectively it is interesting to try to identify the common features of journals in a sub-cluster. In the psychology cluster, A might be called 'general psychology'; B includes all the German psychology journals; C is perhaps 'cure and prevention of psychological disturbance'; and D is too small to merit a name but is close to C. In the economics cluster: A seems general; B is clearly 'political and social economics'; and C and D are biassed towards 'economic statistics and statistical methods'. Even in the amorphous cluster subjective classification reveals some structure: A is probably 'social anthropology'; B 'sociology'. C is clearly related to the psychology group C* and is to do with curing and preventing psychological and social disturbances; D is rather political; F is definitely French; and E and G might be thought of as less academic journals.

* In fact, reference to the original clustering shows that psychology groups C and D and group C from amorphous cluster, together with J. Experimental Psychology, constitute cluster 4 at level 10.

These results are typical of those obtained from other runs of the cluster program. Some parts of the data, in this case psychology and economics, have a well defined structure. Other areas are grey and amorphous. In these it is difficult to decide which of the observed results are true properties of the subject areas and which are peculiarties of the data or how it has been processed.

Another way of looking at the clustering results is to start from the point of view of the clusters rather than that of the journals in them. An attempt can also be made to follow the clusters obtained at one level, as the number of clusters is reduced. As an example the clustering into 8 clusters can be briefly described as follows.

Table 7

Cluster characteristics and description

| Cluster no. | Size | Mean distance from cluster centre | Description |
|---|---|---|---|
| 1 | 25 | 5.6 | Sociology & Social Psychiatry |
| 2 | 4 | 3.3 | French Economics |
| 3 | 19 | 4.8 | Social Anthropology |
| 4 | 8 | 4.2 | Politics |
| 5 | 25 | 3.3 | General Psychology |
| 6 | 10 | 3.7 | Economic Statistics |
| 7 | 12 | 3.7 | Psychology (German plus oddments) |
| 8 | 12 | 2.9 | Economics |

The mean distance from the cluster centre of the items in the cluster is a measure of the compactness of the cluster. Thus cluster 5 is much more compact than cluster 1 even though the number of items in each is the same.

The general psychology cluster 5 is fairly isolated and remains practically unchanged until finally merged with 7, which is also relatively isolated but much less dense. The economics clusters 6 and 8 merge to form a slightly less isolated cluster 6, with a few points lying between it and cluster 1. There is some overlap between clusters 1 and 3, which eventually merge to form a large cluster which is not very compact.

Although the mathematical and computational problems of allowing overlapping clusters are considerable, it is worth considering the output from the program which might be relevant to lifting this restriction. For each clustering the distance of each point from its own cluster, its distance from the next nearest cluster and the ratio of these distances are output. Looking at the three final clusters obtained and if we consider all journals with the ratio

$$\frac{\text{distance to next nearest cluster}}{\text{distance to own cluster centre}} > 1.15$$

as points overlap, we get slightly over 10 per cent of points in more than one cluster. Some of these points can genuinely be considered as belonging to two clusters while others do not appear to fit at all well into either. If it were required to determine which journals should be covered by two secondary services it would clearly be necessary to eliminate the latter type, either by subjective judgement or by some other objective criterion. For instance, it might be required that the distance of a genuine overlap point from its secondary cluster centre be less than one and a half times the mean distance of points in that cluster from the centre.

4.1.4 With cells divided by row totals and a constant added to each cell

Behind this adjustment is the idea that for a journal to be cited at all by a source journal could be considered as more important, in differentiating between it and another journal not cited by the source, than a small difference in the percentage of citations to each of the journals coming from a particular source. It was hoped that this process might reveal more structure in the amorphous cluster. This was not in fact borne out, and the clusterings obtained from 10 to 3 clusters were broadly similar to the results without the addition of the constant (regarding the values of the variables for a journal as the percentage of citations to that journal which came from each source, 10 per cent was added to each value). Since there is no best clustering it is difficult to decide which of two is to be preferred when neither has any obvious relative

merit or disadvantage. Although the detailed composition of the clusters obtained differs slightly, the general description of the clusters revealed in this case and in section 4.1.3 is the same. At the 3-cluster level only four journals were in different clusters.

### 4.1.5 Adjustments to compensate for self-citation

Bias due to self-citation is a well known phenomenon in citation studies. Two methods of compensation were tried. The first was to reduce self-citations by 25 per cent before dividing by row totals. The effect of this adjustment was negligible at all levels between 10 and 4 clusters, and the clusters at the 3-cluster level were identical with those described in section 4.1.3.

Because the extent of self-citation varies considerably between journals this first approach has been criticized. A different approach which has maximum effect on journals with a high rate of self-citation was therefore tried. This was to reduce self-citations which were the highest in their row to the level of the next highest in the row before dividing by row totals. The effect was small but mainly on those source journals which had a high rate of self-citation. This correction was used for all subsequent clustering.

## 4.2 Results using other approaches

To compare results given by the SCICON algorithm with those of some other approaches, the data divided by row totals (Section 4.1.3) for the 34 journals in the psychology cluster, together with American Psychologist and Science, was used with three other algorithms. Those algorithms were chosen as being simple to apply without excessive programming effort. The algorithm suggested by C.J. Van Rijsbergen (1970) appeared as a FORTRAN program in Computer Journal and required only the incorporation of calculation of a similarity matrix. Euclidean distance was used for this purpose. Intermediate output from this program included a listing of the similarity elements sorted in increasing magnitude. This listing was used for hand-calculation of the single - and complete-linkage clusterings.

Van Rijsbergen's algorithm identifies all clusters having the property that the maximum distance between any pair of points in the cluster is less than the minimum distance of any point in the cluster to any point not in the cluster. Twelve small clusters which satisfy this very stringent condition were obtained. Of these, three are inevitable since the variable values for

the journals in each cluster are identical. The nature of the condition for a cluster ensures that any data is bound to contain at least one pair of points satisfying the condition, and not much importance can be attached to two-point clusters. These results merely show that this simple but stringent condition is of no use to DISISS since the data does not form very distinct isolated clusters.

The dendrograms obtained from the single- and complete-linkage algorithms are displayed in Figures 3 and 4. The circled groups identify the tight clusters found by Van Rijsbergen's algorithm. Also added to the figures are the trace codes obtained for the journals in the psychology cluster in section 4.1.3 (see Figure 1). The following facts should be noted when considering the dendrograms.

(i) The exact ordering of the journals is irrelevant. The hierarchy could have been displayed using a large number of different permutations of the journals. For instance the sequence of the journals in Figure 6 between J. Abnormal and Social Psychology and Psychologische Forschung could have been placed before Acta Psychologica.

(ii) In Figure 10 (single-linkage) the level* at which two sub-clusters combine to form one cluster is the minimum distance between any point in one sub-cluster and any point in the other.

(iii) In Figure 11 (complete-linkage) the level at which two sub-clusters combine is the maximum distance between any pair of points in the combined cluster.

Both methods clearly identify the journals with trace code B (and J. Psychology, which lay between A and B) as a cluster at a relatively low level, and at a similar level a large proportion of the A journals are clustered. The exact behaviour of the more peripheral (in terms of clustering rather than of subject matter) journals differs, the difference being a function of the particular algorithms. The complete-linkage dendrogram also identifies the C and D groups reasonably well and produces the type of clustering which could be of use to DISISS. The 'chaining' tendency of the single-linkage algorithm tends to obscure the structure of Figure 1, which could be very useful in the design of secondary services.

* The use of the term 'level' in this context should not be confused with the level of clustering referred to with the SCICON method, which is the number of clusters.

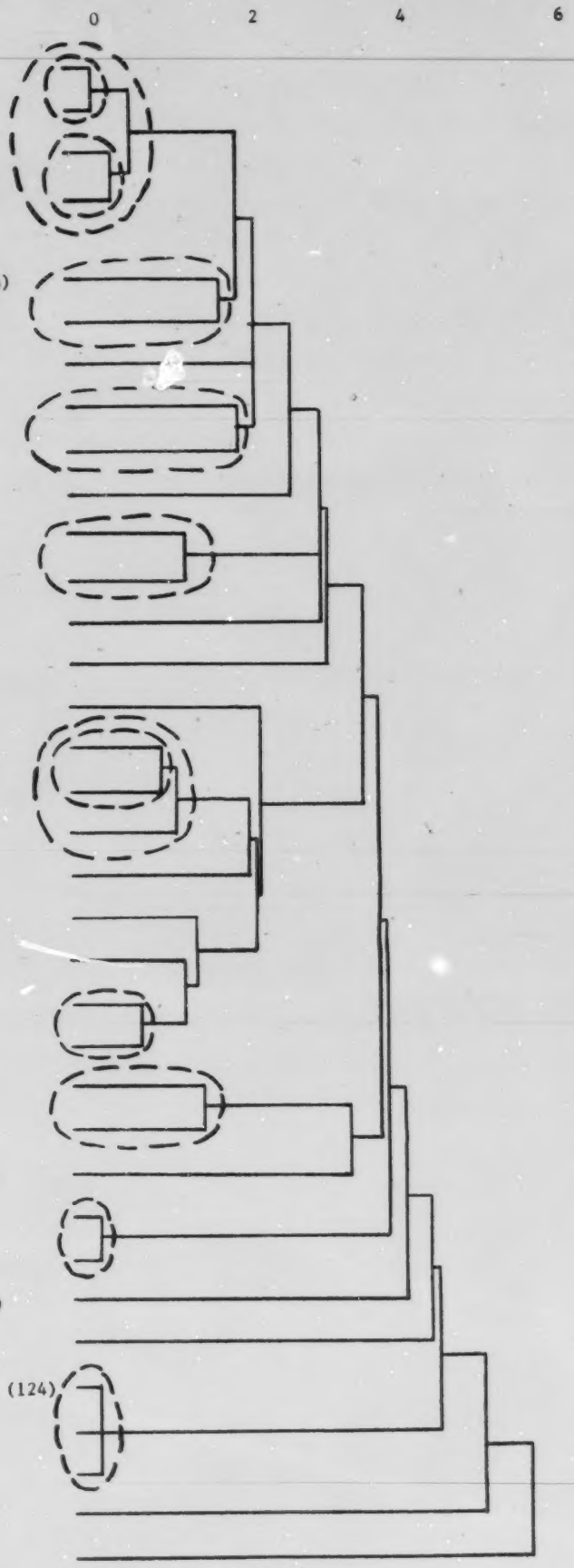

0
2
4
6
A 1 ACTA PSYCHOLOGICA (002)
A 7 ARCHIVES DE PSYCHOLOGIE (032)
A 22 J.NEUROPHYSIOLOGY (125)
A 2 AM.J. PSYCHOLOGY (014)
A 32 PSYCHOLOGICAL MONOGRAPHS (176)
A 18 J. COMPARATIVE AND PHYSL. PSYCH (114)
A 11 BR.J. PSYCHOLOGY (048)
A 27 NATURE (154)
A 20 J.GENETIC PSYCHOLOGY (119)
A 13 CANADIAN J.PSYCHOLOGY (060)
A 31 PSYCHOLOGICAL BULLETIN (175)
A 28 OCCUPATIONAL PSYCHOLOGY (159)
A 12 BULL.BR. PSYCHOLOGICAL SOC. (051)
A 14 GENETIC PSYCHOLOGY (092)
A 25 J.SOCIAL PSYCHOLOGY (130)
B 16 J.ABNORMAL AND SOCIAL PSYCH. (109)
B 33 PSYCHOLOGICAL REVIEW (178)
B 29 PHILOSOPHICAL STUDIES (166)
B 23 J.PERSONALITY (126)
24 J.PSYCHOLOGY (129)
B 30 PSYCHOLOGISCHE ARBEITEN (173)
B 36 STUDIUM GEN. (215)
B 6 ARCHIV FUR GESAMTE PSYCH. (028)
B 34 PSYCHOLOGISCHE FORSCHUNG (180)
3 AMERICAN PSYCHOLOGIST (018)
D 19 J.EDUCATIONAL PSYCH. (116)
D 17 J.COMP. PSYCH (113)
A 15 HARVARD LAW REVIEW (098)
A 5 ANNALS N.Y. ACAD. SCIENCE (024)
26 J.OPERATIONS RESEARCH SOC. AM. (136)
35 SCIENCE (202)
C 21 J.NEUROLOGY, NEUROSURGERY AND PSYCH (124)
C 10 BR. J. PSYCHIATRY (047)
C 9 BR.J. MEDICAL PSYCH (046)
C 8 BR.J. EDUC. PSYCH (045)
4 ANNALS MATH. STAT (022)

Figure 3

Single-linkage hierarchical clustering of psychology cluster

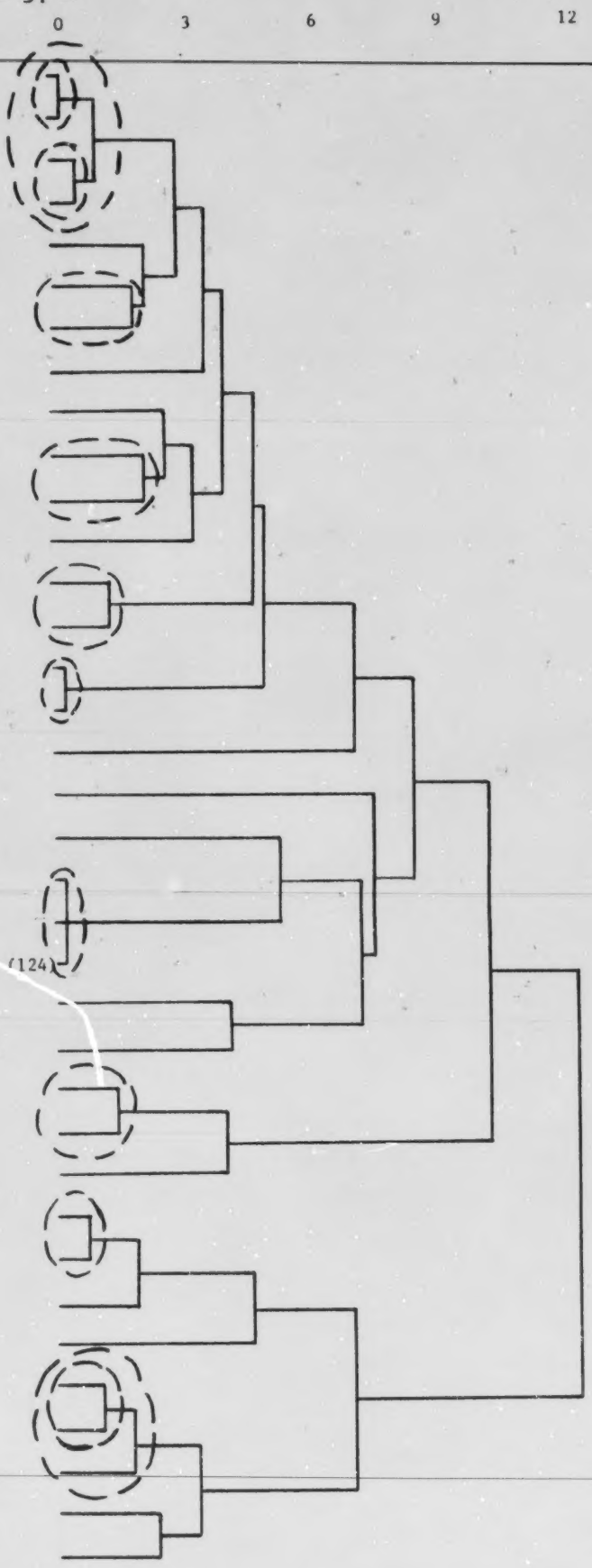

0
3
6
9
12
A 1 ACTA PSYCHOLOGICA (002)
A 7 ARCHIVES DE PSYCHOLOGIE (032)
A 22 J. NEUROPHYSIOLOGY (125)
A 2 AM.J. PSYCHOLOGY (014)
A 32 PSYCHOLOGICAL MONOGRAPHS (176)
A 18 J.COMPARATIVE AND PHYSL.PSYCH (114)
A 11 BR.J. PSYCHOLOGY (048)
A 14 GENETIC PSYCHOLOGY (092)
A 27 NATURE (154)
A 20 J.GENETIC PSYCHOLOGY (119)
A 13 CANADIAN J. PSYCHOLOGY (060)
A 31 PSYCHOLOGICAL BULLETIN (175)
A 28 OCCUPATIONAL PSYCHOLOGY (159)
A 12 BULL.BR.PSYCHOLOGICAL SOC.(051)
A 5 ANNALS N.Y.ACAD SCIENCE (024)
A 15 HARVARD LAW REVIEW (098)
4 ANNALS MATH. STAT. (022)
35 SCIENCE (202)
C 8 BR.J.EDUC.PSYCH. (045)
C 9 BR.J.MEDICAL PSYCH. (046)
C 10 BR.J.PSYCHIATRY (047)
C 21 J.NEUROLOGY, NEUROSURGERY AND PSYCH. (124)
A 25 J.SOCIAL PSYCHOLOGY (130)
26 J.OPERATIONS RESEARCH SOC. AM. (136)
3 AMERICAN PSYCHOLOGIST (018)
D 19 J.EDUCATIONAL PSYCH. (116)
D 17 J.COMPARATIVE PSYCH. (113)
B 6 ARCHIV FOR GESAMTE PSYCH. (028)
B 34 PSYCHOLOGISCHE FORSCHUNG (180)
B 36 STUDIUM GEN. (215)
B 30 PSYCHOLOGISCHE ARBEITEN (173)
B 29 PHILOSOPHICAL STUDIES (166)
B 33 PSYCHOLOGICAL REVIEW (178)
B 23 J.PERSONALITY (126)
B 16 J.ABNORMAL AND SOCIAL PSYCH. (109)
24 J.PSYCHOLOGY (129)

Figure 4

Complete-linkage hierarchical clustering of psychology cluster

## 4.3 Conclusions from the Pilot Study

(i) Since the single-and complete-linkage algorithms represent extremes of their family of hierarchical methods, it is encouraging that they and the SCICON method identify the B clusters as distinct from the larger A group, and less importantly, detect the C and D groups. This implies that the structure revealed by the SCICON method is in the data and not imposed by the algorithm.

(ii) Van Rijsbergen's algorithm and single linkage can be discarded as not providing results of use to DISISS.

(iii) The complete-linkage algorithm provides useful results but is computationally impracticable for large amounts of data.

(iv) The most suitable treatment for the raw data appears to be to divide data cells by their row totals after reducing self-citation cells which are the highest in their row to the level of the next highest cell. Apart from self-citations this means that the data used is the percentage of citations to a journal which are found in each of the source journals. The possibility of adding a constant to each non-zero cell should not however be discarded.

(v) The source journals which appeared as cited titles but were cited only by themselves were included in the clustering runs but tended to be outliers distorting the clusters in the rest of the data. The SCICON method is unlikely to produce clusters of very different sizes and thus outliers will not appear as single-point clusters. For this reason it is probably better to exclude all titles cited by only one source.

## 5.0 THE CRIMINOLOGY CLUSTERS

The three sets of criminology data, citations from articles written in 1950, 1960 and 1970, were clustered separately. Each set of data was normalized as described in Section 4.3, that is all journals cited only once were discarded; self-citations were adjusted to equal the most citations to the cited journal from any other journal; and the observations were scaled by dividing by the total number of citations to that journal (row totals). The numbers of cited journals actually clustered were as follows.

Table 8

Number of Criminology Journals Clustered

| Sample Year | Number of Journals |
|---|---|
| 1950 | 18 |
| 1960 | 49 |
| 1970 | 108 |

The number of citations to many of these journals was very small. Correspondingly the clusters need to be interpreted with care. In many instances one more or one less citation could transfer a journal to a neighbouring cluster.

### 5.1 1950 Sample

Since the number of cited journals in the 1950 sample and the number of citations to each was small, the maximum number of clusters used was five. The results are illustrated in Figure 5. The eighteen journals appear to divide into three main groups plus three journals, Harvard Law Review, Lancet and the British Medical Journal. For what they are worth, the three groups and their distinguishing features are as follows.

A. Journals cited by the Journal of Criminal Law, Criminology and Police Science and usually by Psychiatry.

| | |
|---|---|
| 348 | Amer. Journal Orthopsychiatry |
| 352 | Amer. Journal Psychiatry |
| 3449 | Journal Clin. Psychopathology |
| 4177 | Journal Crim. Law, Criminology and Pol. Sci. |
| 1568 | Journal of Psychology |
| 4383 | Teachers College Record |

B. Journals cited by the Journal of Social Issues.

| | |
|---|---|
| 1383 | Human Relations |
| 1458 | International Journal of Psychoanalysis |
| 3396 | Journal Abnormal and Soc. Psychology |
| 2237 | Scientific American |
| 2262 | Social Forces |

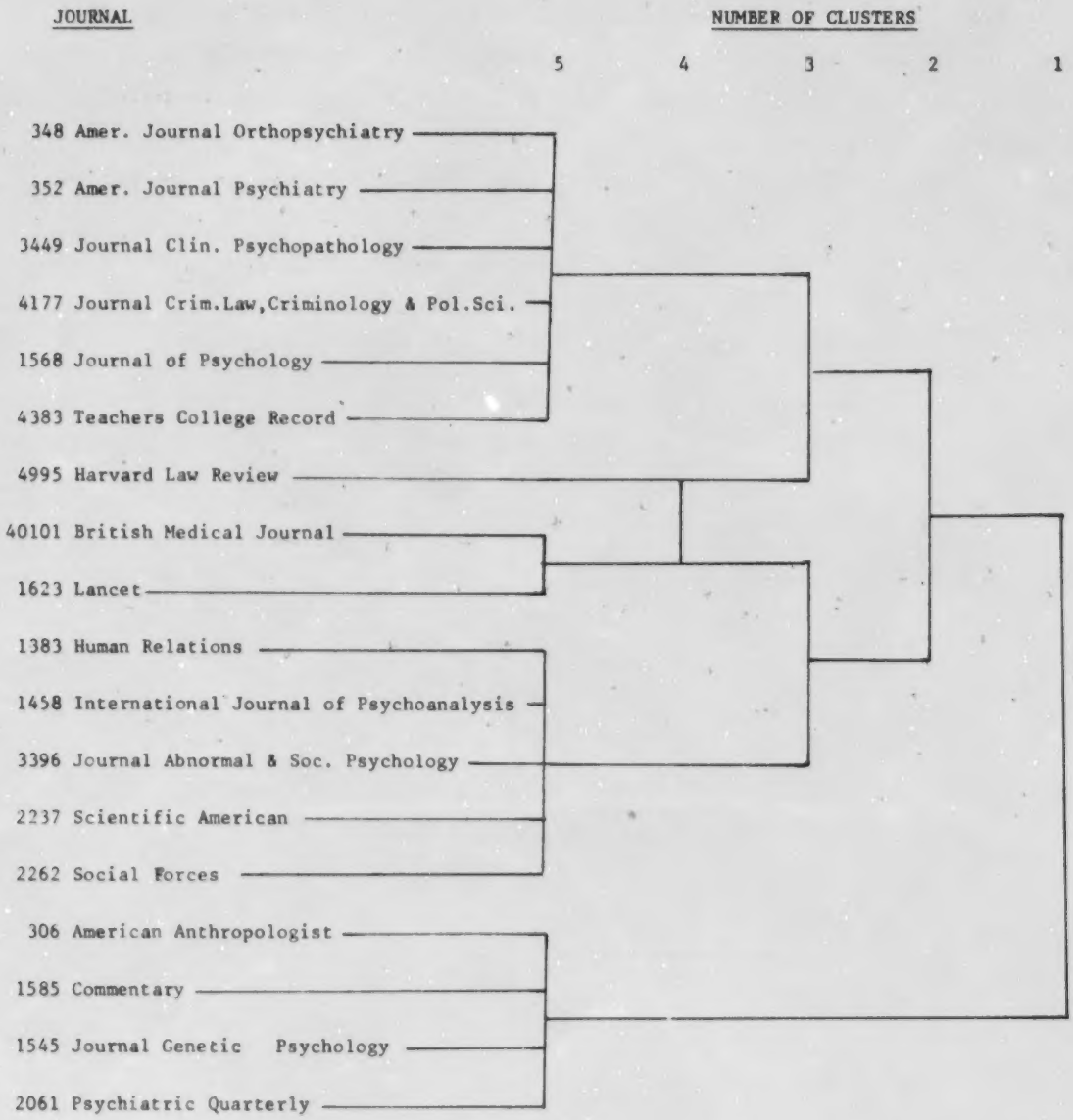
C. Journals which are cited by both Psychiatry and American Sociological Review.

| | |
|---|---|
| 306 | American Anthropologist |
| 1585 | Commentary |
| 1545 | Journal Genetical Psychology |
| 2061 | Psychiatric Quarterly |

There is so little data in the 1950 sample that these results must be treated merely as a curiosity.

Figure 13

Criminology clusters (1950)



5.2 1960 Sample

The 1960 sample shows much more structure. The clusters are shown in Figure 6. In examining them it is important to notice how the clusters merge and split as the number is reduced. Parts of Figure 6 show crisp hierarchical structuring which indicates a definite gap between clusters. Other parts show clusters splitting rather than merging; this indicates a grey area which is being artificially clustered by the program and would probably be divided differently by another computer algorithm. The program was run for numbers of clusters up to 15. Two large groups of journals stick tightly together. They are as follows.

I A group of nine journals which could be described as medical psychology.

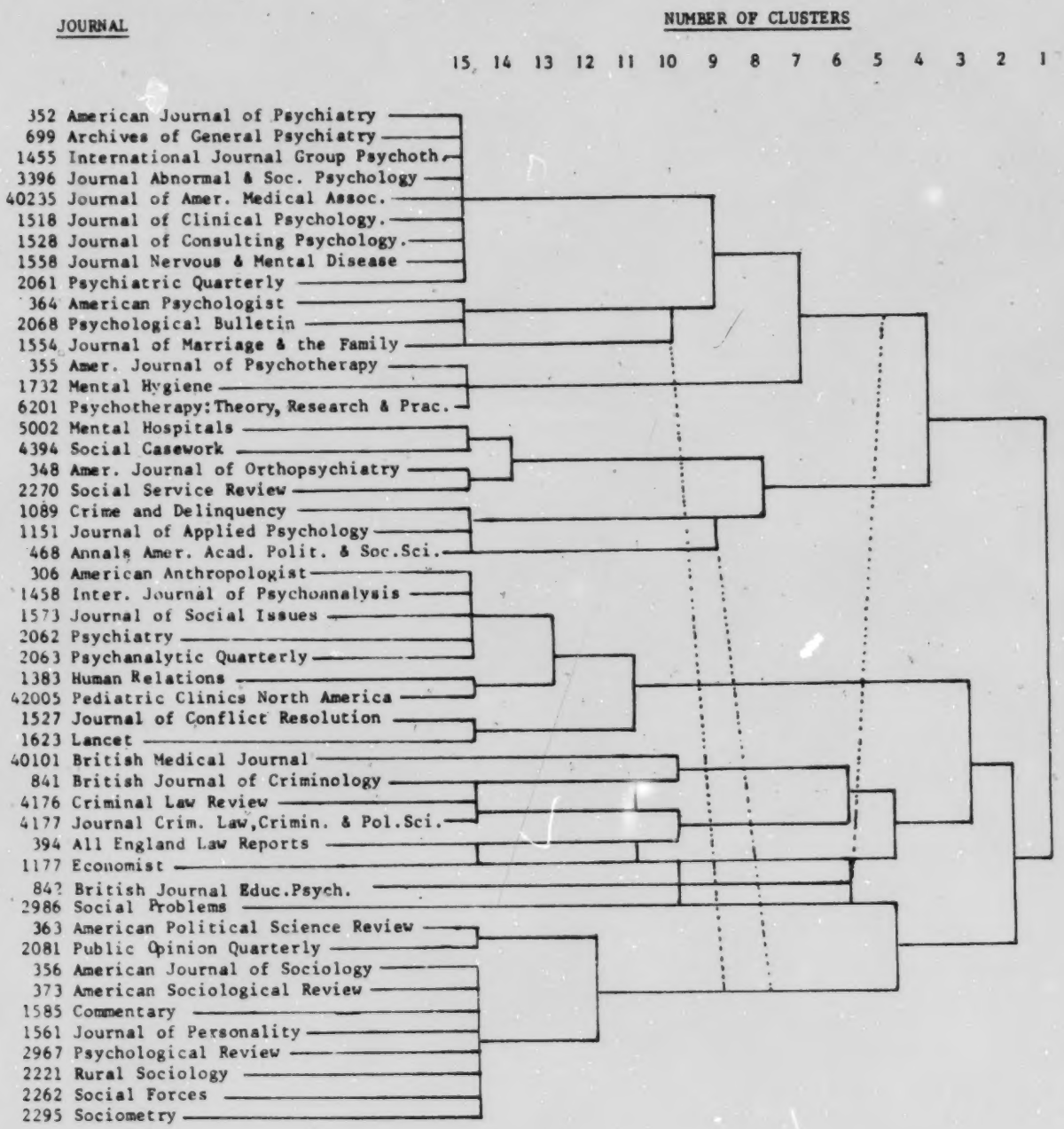
| | |
|---|---|
| 352 | American Journal of Psychiatry |
| 699 | Archives of General Psychiatry |
| 1455 | International Journal Group Psychotherapy |
| 3396 | Journal Abnormal and Soc. Psychology |
| 40235 | Journal of Amer. Medical Assoc. |
| 1518 | Journal of Clinical Psychology |
| 1528 | Journal of Consulting Psychology |
| 1558 | Journal Nervous and Mental Disease |
| 2061 | Psychiatric Quarterly |

II A group of eight mainly sociology journals.

| | |
|---|---|
| 365 | American Journal of Sociology |
| 373 | American Sociological Review |
| 1585 | Commentary |
| 1561 | Journal of Personality |
| 2967 | Psychological Review |
| 2221 | Rural Sociology |
| 2262 | Social Forces |
| 2295 | Sociometry |

Figure 6

Criminology clusters (1960)


JOURNAL
NUMBER OF CLUSTERS
15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
352 American Journal of Psychiatry
699 Archives of General Psychiatry
1455 International Journal Group Psychoth.
3396 Journal Abnormal & Soc. Psychology
40235 Journal of Amer. Medical Assoc.
1518 Journal of Clinical Psychology.
1528 Journal of Consulting Psychology.
1558 Journal Nervous & Mental Disease
2061 Psychiatric Quarterly
364 American Psychologist
2068 Psychological Bulletin
1554 Journal of Marriage & the Family
355 Amer. Journal of Psychotherapy
1732 Mental Hygiene
6201 Psychotherapy: Theory, Research & Prac.
5002 Mental Hospitals
4394 Social Casework
348 Amer. Journal of Orthopsychiatry
2270 Social Service Review
1089 Crime and Delinquency
1151 Journal of Applied Psychology
468 Annals Amer. Acad. Polit. & Soc.Sci.
306 American Anthropologist
1458 Inter. Journal of Psychoanalysis
1573 Journal of Social Issues
2062 Psychiatry
2063 Psychanalytic Quarterly
1383 Human Relations
42005 Pediatric Clinics North America
1527 Journal of Conflict Resolution
1623 Lancet
40101 British Medical Journal
841 British Journal of Criminology
4176 Criminal Law Review
4177 Journal Crim. Law, Crimin. & Pol.Sci.
394 All England Law Reports
1177 Economist
842 British Journal Educ.Psych.
2986 Social Problems
363 American Political Science Review
2081 Public Opinion Quarterly
356 American Journal of Sociology
373 American Sociological Review
1585 Commentary
1561 Journal of Personality
2967 Psychological Review
2221 Rural Sociology
2262 Social Forces
2295 Sociometry

For less than 4 clusters only the hierarchy is shown, not individual exceptions.

Two other journals are associated with this group except at levels of thirteen or more clusters.

363 American Political Science Review
2081 Public Opinion Quarterly

III The following five heavily cited journals also keep together to form a law and criminology cluster.

40101 British Medical Journal
841 British Journal of Criminology
4176 Criminal Law Review
4177 Journal Crim. Law, Crimin. and Pol. Sci.
394 All England Law Reports

IV Another clear cut group is three journals on psychotherapy.

355 Amer. Journal of Psychotherapy
6201 Psychotherapy: Theory, Research and Practice
1732 Mental Hygiene

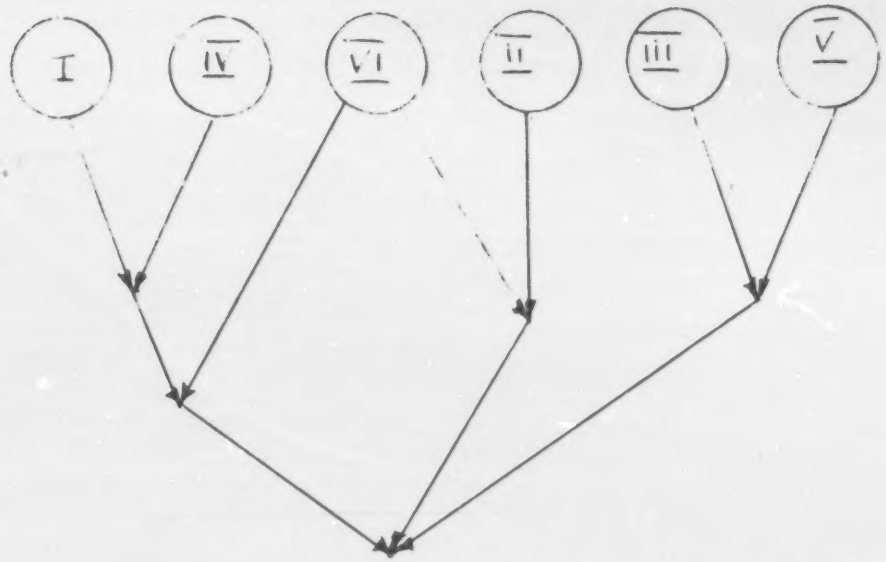
V The final group that the algorithm separates out is as follows. Many of them are lightly cited, with the majority of citations from Psychiatry. It is probable that this group is a chance result of the data and clustering algorithm used.

1623 Lancet
1527 Journal of Conflict Resolution
42005 Pediatric Clinics North America
1383 Human Relations
2063 Psychoanalytic Quarterly
2062 Psychiatry
306 American Anthropologist
1458 Inter. Journal of Psychoanalysis
1573 Journal of Social Issues

VI As Figure 6 shows, most of the other journals belong to no obvious grouping, but seven journals form a loose grouping that appears to be more solid than merely noise in the data.

| | |
|---|---|
| 468 | Annals Amer. Acad. Polit. and Soc. Sci. |
| 1089 | Crime and Delinquency |
| 1151 | Journal of Applied Psychology |
| 2270 | Social Service Review |
| 348 | Amer. Journal of Orthopsychiatry |
| 5002 | Mental Hospitals |
| 4394 | Social Casework |

The hierarchical relationship between these main groupings is shown below.



## 5.3 1970 Sample

The 1970 sample is much larger both in the number of cited journals and in the number of citations to each. Consequently there are more journals that do not fit into any clear cut groupings. However if a certain amount of poet's licence is allowed much of the same structure can be seen as in the 1960 data. The results are shown in Figure 7. Since 108 journals are too many to show in a single diagram, journals that always appear in the same cluster have been grouped together and identified by a letter. The following letters have been used.

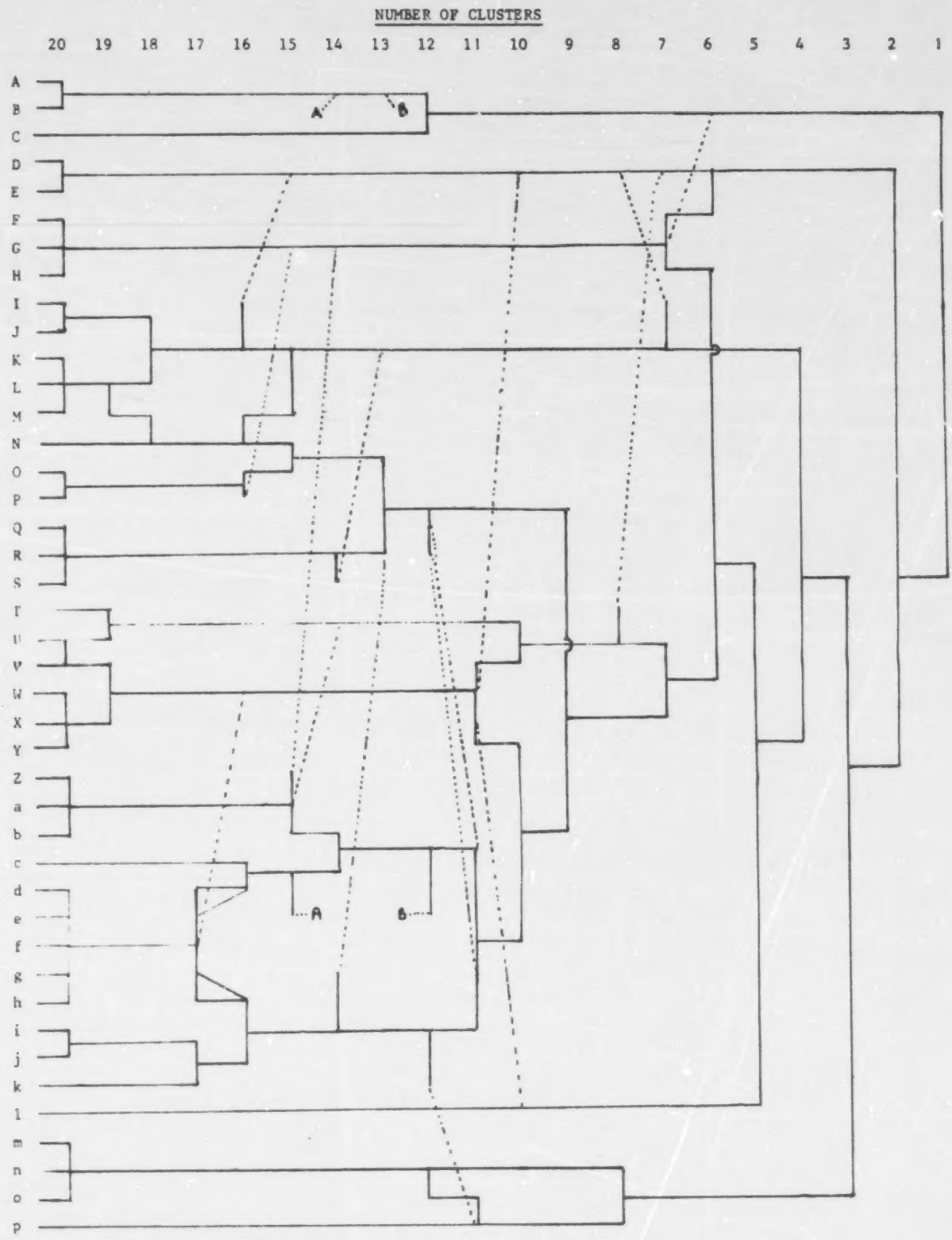
| Group | Size | Journals | |
|---|---|---|---|
| A | 1 | 858 | British Journal of Psychiatry |
| B | 7 | 40047 | Annals of Internal Medicine |
| | | 1269 | Family Process |
| | | 40235 | Journal of the American Medical Association |
| | | 40238 | Journal of Applied Physiology |
| | | 42004 | Journal of Pharmacology and Experimental Therapeutics |
| | | 40355 | New England Journal of Medicine |
| | | 2062 | Psychiatry |
| C | 27 | 131 | Acta Psychiatrica Scandinavia |
| | | 135 | Acta Psychologica |
| | | 40024 | American Journal of Medicine |
| | | 352 | American Journal of Psychiatry |
| | | 355 | American Journal of Psychotherapy |
| | | 357 | American Journal of Public Health |
| | | 699 | Archives of General Psychiatry |
| | | 687 | Archiv für Psychiatrie und Nervenkrankheiten |
| | | 805 | Behaviour Research and Therapy |
| | | 859 | British Journal of Psychology |
| | | 861 | British Journal of Social and Clinical Psychology |
| | | 863 | British Journal of Social Psychiatry |
| | | 851 | British Journal of Statistical Psychology |
| | | 42003 | Bulletin on Narcotics |
| | | 40158 | Diseases of the Nervous System |
| | | 40163 | Electroencephalography and Clinical Neurophysiology |
| | | 40207 | Human Biology |
| | | 4997 | Internal Journal of the Addictions |
| | | 3996 | Journal of Abnormal and Social Psychology |
| | | 1558 | Journal of Nervous and Mental Disease |

| Group | Size | Journal | |
|---|---|---|---|
| C | 27 | 1568 | Journal of Psychology |
| | | 1569 | Journal of Psychosomatic Research |
| | | 1623 | Lancet |
| | | 1878 | Nature |
| | | 2061 | Psychiatric Quarterly |
| | | 2968 | Psychosomatic Medicine |
| | | 5254 | Quarterly Journal of Studies on Alcohol |
| D | 6 | 348 | American Journal of Orthopsychiatry |
| | | 1528 | Journal of Consulting Psychology |
| | | 1562 | Journal of Personality and Social Psychology |
| | | 2064 | Psychoanalytic Study of the Child |
| | | 2069 | Psychological Issues |
| | | 2237 | Scientific American |
| E | 3 | 988 | Child Development |
| | | 42004 | Comprehensive Psychiatry |
| | | 2234 | Science |
| F | 1 | 306 | American Anthropologist |
| G | 4 | 1458 | International Journal of Psychoanalysis |
| | | 5083 | Journal of Health and Social Behaviour |
| | | 2063 | Psychoanalytic Quarterly |
| | | 2966 | Psychological Monographs |
| H | 1 | 3449 | Journal of Clinical Psychopathology |
| I | 1 | 1518 | Journal of Clinical Psychology |
| J | 2 | 804 | Behavioral Science |
| | | 1459 | International Journal of Psychiatry |
| K | 1 | 1898 | New Republic |
| L | 4 | 444 | Annales Internationales Criminologie |
| | | 706 | Archivio di Psichologia, Neurologia e Psichiatria |
| | | 4958 | Federal Probation |
| | 4 | 6203 | Revue de Science Criminelle et de Droit Pénal Comparé |
| M | 1 | 42001 | Annales Medicine Légale |
| N | 2 | 394 | All England Law Reports |
| | | 4176 | Criminal Law Review |
| O | 1 | 1601 | Juridical Review |
| P | 1 | 1638 | Law and Contemporary Problems |
| Q | 1 | 5848 | Journal of Forensic Sciences |
| R | 1 | 42007 | International Criminial Police Review |
| S | 1 | 5661 | Kriminalistik |

| Group | Size | Journal | |
|---|---|---|---|
| T | 3 | 1732 | Mental Hygiene |
| | | 4394 | Social Casework |
| | | 4187 | Social Work |
| U | 1 | 4887 | Child Welfare |
| V | 1 | 5291 | Saturday Review |
| W | 1 | 5212 | Perceptual and Motor Skills |
| X | 1 | 40049 | Annals of New York Academy of Science |
| Y | 1 | 1462 | International Journal of Social Psychiatry |
| Z | 1 | 2065 | Psychoanalytic Review |
| a | 1 | 5092 | Journal of Jewish Communal Service |
| b | 2 | 1089 | Crime and Delinquency |
| | | 6201 | Psychotherapy: Theory, Research and Practice |
| c | 3 | 42002 | Annals (Journal of Natural History) |
| | | 5110 | Journal of Research in Crime and Delinquency |
| | | 5394 | Yale Law Journal |
| d | 1 | 1575 | Journal of Social Psychology |
| e | 1 | 1901 | New Society |
| f | 1 | 468 | Annals American Academy of Political and Social Science |
| g | 1 | 841 | British Journal of Criminology |
| h | 1 | 40101 | British Medical Journal |
| i | 3 | 4177 | Journal of Criminal Law, Criminology and Police Science |
| | | 2068 | Psychological Bulletin |
| | | 2986 | Social Problems |
| j | 2 | 364 | American Psychologist |
| | | 5244 | Psychology Today |
| k | 2 | 2010 | Phylon |
| | | 2386 | Transaction |
| l | 4 | 4995 | Harvard Law Review |
| | | 5148 | Minnesota Law Review |
| | | 5865 | Texas Law Review |
| | | 5862 | University of Pennsylvannia Law Review |
| m | 1 | 815 | Biometrics |
| n | 3 | 161 | Administrative Science Quarterly |
| | | 5072 | Journal of Educational Sociology |
| | | 5251 | Public Interest |
| o | 1 | 356 | American Journal of Sociology |
| p | 6 | 363 | American Political Science Review |
| | | 373 | American Sociological Review |
| | | 1577 | Journal of the American Statistical Association |
| | | 1554 | Journal of Marriage and the Family |

Figure 7

Criminology clusters (1970)




For less than 5 clusters only the hierarchy is shown, not individual exceptions.

| Group | Size | Journal | |
|---|---|---|---|
| p | 6 | 1573 | Journal of Social Issues |
| | | 4496 | Pacific Sociological Review |

I As with the 1960 data, the most distinct group consists of 34 journals in medical psychology.

7 journals in section B
27 journals in section C

II Also reflecting the structure noted in the 1960 data are 11 journals mainly in sociology.

1 journal in section m
3 journals in section n
1 journal in section o
6 journals in section p

III Four journals form a tight academic law cluster.

4 journals in section l

IV Seventeen journals form another tight cluster. They are mainly in psychotherapy.

6 journals in section D
3 journals in section E
1 journal in section F
4 journals in section G
1 journal in section H
1 journal in section P
1 journal in section Z

V Six journals form the final tight cluster in social work.

3 journals in section T
1 journal in section U
1 journal in section V
1 journal in section a

The remaining journals do not form such clear cut clusters. They form a large overlapping group in which the borders between clusters are wide grey areas. However within this big grouping the following features are worth noting.

VI A grey area of criminology journals, some or all of which appear together at most levels of clustering.

2 journals in section b
3 journals in section c
1 journal in section d
1 journal in section e
1 journal in section g
3 journals in section i
2 journals in section j

VII A grey area of police and law journals, some or all of which appear together at most levels of clustering.

2 journals in section N
1 journal in section O
1 journal in section Q
1 journal in section R
1 journal in section S

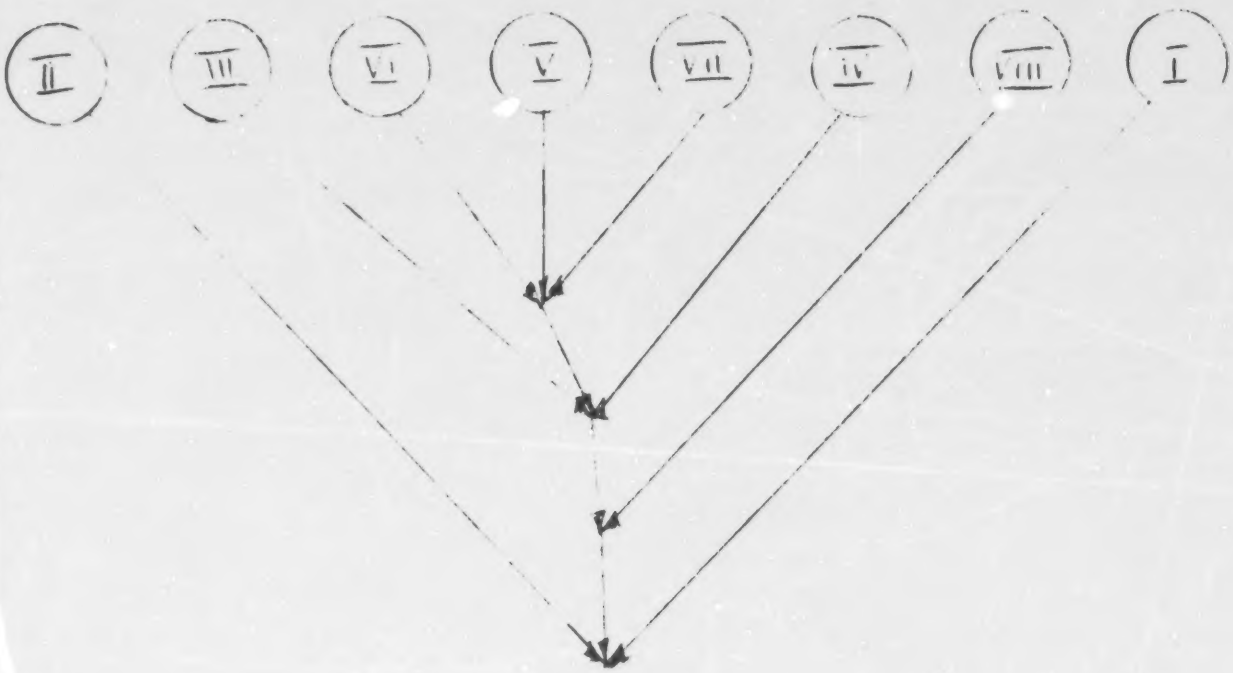
VIII Finally there are two grey areas which overlap each other. The first is an area of clinical psychology.

1 journal in section I
2 journals in section J
1 journal in section K

The other area is mainly foreign journals.

4 journals in section L

These eight groups and grey areas have the following hierarchical relationship.



## 5.4 Changes between 1960 and 1970

When the small size of the 1960 sample is taken into consideration the main groupings in 1960 and 1970 appear to be roughly similar. The following thirty-five journals appear in both the samples.

Table 9

Criminology journals in 1960 and 1970 clusters

| | Journal | 1960 Group | 1970 Group |
|---|---|---|---|
| 1623 | Lancet | cited by Psychiatry | medical psychology |
| 2063 | Psychoanalytic Quarterly | cited by Psychiatry | psychotherapy |
| 2062 | Psychiatry | cited by Psychiatry | medical psychology |
| 306 | American Anthropologist | cited by Psychiatry | psychotherapy |
| 1458 | International Journal of Psychoanalysis | cited by Psychiatry | psychotherapy |
| 1573 | Journal of Social Issues | cited by Psychiatry | sociology |
| 40101 | British Medical Journal | law and criminology | - |
| 841 | British Journal of Criminology | law and criminology | criminology |
| 4176 | Criminal Law Review | law and criminology | criminology |
| 4177 | Journal of Criminal Law, Criminology and Police Science | law and criminology | police and law |
| 394 | All England Law Reports | law and criminology | police and law |
| 2986 | Social Problems | - | criminology |
| 363 | American Political Science Review | sociology | sociology |
| 356 | American Journal of Sociology | sociology | sociology |
| 373 | American Sociological Review | sociology | sociology |
| 2967 | Psychological Review | sociology | psychotherapy |
| 468 | Annals Amer. Acad. Polit. and Soc. Sci. | loose group | - |
| 1089 | Crime and Delinquency | loose group | criminology |
| 1151 | Journal of Applied Psychology | loose group | medical psychology |
| 348 | American Journal of Orthopsychiatry | loose group | psychotherapy |
| 4394 | Social Casework | loose group | social work |
| 1554 | Journal of Marriage and the Family | - | sociology |
| 364 | American Psychologist | - | criminology |
| 2068 | Psychological Bulletin | - | criminology |
| 352 | American Journal of Psychiatry | medical psychology | medical psychology |
| 699 | Archives of General Psychiatry | medical psychology | medical psychology |
| 3396 | Journal Abnormal and Soc. Psychology | medical psychology | medical psychology |
| 40235 | Journal of the American Medical Association | medical psychology | medical psychology |
| 1518 | Journal of Clinical Psychology | medical psychology | clinical psychology |
| 1528 | Journal of Consulting Psychology | medical psychology | medical psychology |
| 1958 | Journal of Nervous and Mental Disease | medical psychology | medical psychology |
| 2061 | Psychiatric Quarterly | medical psychology | medical psychology |
| 355 | American Journal of Psychotherapy | psychotherapy | medical psychology |
| 6201 | Psychotherapy: Theory, Research and Practice | psychotherapy | criminology |
| 1732 | Mental Hygiene | psychotherapy | social work |

Each of the four main groups found in 1960 more or less matches one or more corresponding groups found in 1970 journals. However the patterns do not match exactly and without further data it is not possible to determine how much of the difference comes from true changes between the two dates and how much from random elements in the data.

## 6.0 CLUSTERS FROM THE MAIN CITATION FILE

The number of journals clustered in the pilot study and from the criminology file was small enough to allow individual journals to be examined. The two sets of data derived from the main citation file were much bigger. Even after removing journals cited by only one source the ranked sample had 495 cited journals and the random sample had 254 journals. The following table gives the overlap between the two samples.

Table 10

Overlap Between Ranked and Random Journal Samples

| | Number of journals | % of ranked sample | % of random sample |
|---|---|---|---|
| In ranked only | 311 | 63 | - |
| In random only | 70 | - | 29 |
| In both | 184 | 37 | 71 |
| Total | 565 | 100 | 100 |

Social science sources cite journals which are not social science journals. Of these 565 journals, all but 110 (19 per cent) had previously been identified as social science journals. The remaining 110 are mainly in medicine and the natural sciences.

### 6.1 Citations from the ranked list of journals

The analysis of the pilot study and criminology data looked at individual journals. The source journals in the ranked and random samples each cited too many journals to look at individually. The journals were clustered at all levels from 30 to 2 but for clarity the analysis has been concentrated on 30, 20 and 10 clusters.

Table 11 gives the entire clustering at the 30 cluster level for journals cited by the ranked list of source journals. Some of these cluters are easy to label. Others are apparently more vague. Table 12 is an attempt to label each of the 30 clusters.

Table II

30 Clusters from Ranked Sample

CLUSTER 1

| | |
|---|---|
| 1319 | Geographical Rev |
| 2008 | Philosophy of Sci |
| 2237 | Scientific American |
| 3886 | Cuadernos Americanos |
| 4172 | Am Antiquity |
| 4812 | Arctic Anthropology |
| 4857 | Bull Bureau Am Ethnology |
| 5007 | Ibero-Americana |
| 5179 | New Mexico |
| 5205 | Papers Peabody Museum |
| 40031 | Am J Sci |
| 40140 | Communicaciones del Instituto Tropical de Investigaciones Cientificas de la Universidad de El Salvador |
| 40425 | Am J Sci Radiocarbon Suppl |
| 40493 | Stain Technology |
| 40515 | Trans Am Philosophical Soc |

CLUSTER 2

| | |
|---|---|
| 1171 | Economie Appliquée |
| 2977 | Recherches Economiques de Louvain |
| 4181 | Planovoe Khoziaistvo |
| 4185 | Revue d'Economie Politique |
| 4870 | Monthly Rev Federal Reserve Bank N. York |
| 5269 | Rivista Internazionale di Scienze Economiche e Commerciale |
| 5299 | Schweizerische Zeitschrift für Volkswirtschaft und Statistik |
| 5583 | Weltwirtschaftliches Archiv |
| 40106 | Bull Am Math Soc |

CLUSTER 3

| | |
|---|---|
| 352 | Am J Psychiatry |
| 357 | Am J Public Health & Nations Health |
| 468 | Annals Am Acad Political & Soc Sci |
| 804 | Behavioral Sci |
| 815 | Biometrics |
| 864 | Brit J Sociol |
| 1383 | Human Relations |
| 1413 | Industrial Relations |
| 1510 | J Applied Behavioral Sci |
| 1561 | J Personality |
| 1573 | J Social Issues |
| 1784 | Monthly Labor Rev |
| 2062 | Psychiatry |
| 2063 | Psychoanalytic Q |
| 2065 | Psychoanalytic Rev |
| 2068 | Psychological Issues |
| 2295 | Sociometry |
| 3396 | J Abnormal & Social Psychol |
| 4421 | Business Week |
| 5083 | J Health & Social Behavior |
| 5087 | J Human Resources |
| 5228 | Proc Am Stat Assoc. |
| 5243 | Psychol in Schools |
| 5251 | Public Interest |
| 40355 | New England J Medicine |
| 41166 | Genetic Studies |
| 41343 | Popular Sci Monthly |

CLUSTER 4

| | |
|---|---|
| 327 | Am Educational Res J |
| 1526 | J Am Psychoanalytic Assoc |
| 1536 | J Educational Res |
| 1568 | J Psychol |
| 1575 | J Social Psychol |

CLUSTER 4 (continued)

| | |
|---|---|
| 5008 | IEEE Trans Human Factors in Electronics |
| 5014 | Indian Psychol Rev |
| 5078 | J Experimental Education |
| 40028 | Am J Optometry |
| 40757 | Pharmacol Rev |
| 41339 | Physiol Zool |
| 41361 | Q Rev Biology |

CLUSTER 5

| | |
|---|---|
| 434 | Animal Behaviour |
| 5038 | Jap Psychol Res |
| 5059 | J Comparative & Physiological Psychol |
| 5406 | Psychol Learning & Motivation |
| 5407 | Psychopharmacologia |
| 6188 | Psychonomic Bull |
| 40049 | Annals NY Acad Sci |
| 40084 | Brain |
| 40085 | Brain Res |
| 40143 | Comptes Rendus Académie des Sciences Paris |
| 40163 | Electroencephalography & Clinical Neurophysiology |
| 40177 | Experimental Neurology |
| 40238 | J Applied Physiology |
| 40252 | J Comparative Neurology |
| 40275 | J Neurophysiology |
| 40280 | J Physiology |
| 40352 | Neurology |
| 40404 | Proc Nat Acad Sci USA |
| 40410 | Proc Soc Exp Biol & Med |
| 40658 | Endeavour |
| 40663 | Experientia |
| 40759 | Physiol Rev |
| 41003 | Acta Endocrinol |
| 41006 | Acta Physiol Scand |
| 41023 | Am J Physiology |
| 41028 | American Zoologist |
| 41054 | Arch Int Pharmacodyn |
| 41088 | Brain Behavior and Evolution |
| 41100 | Brain Function |
| 41124 | Communications Behav |
| 41127 | Conditional Reflex |
| 41153 | Exp Brain Res |
| 41222 | J Endocrinology |
| 41256 | J Pharmacology & Experimental Therapeutics |
| 41283 | Life Sciences |
| 41342 | Physiology and Behavior |

CLUSTER 6

| | |
|---|---|
| 203 | America Indigena |
| 314 | Am Behavioral Sci |
| 1268 | Family Process |
| 1462 | Int J Social Psychiatry |
| 1732 | Mental Hygiene |
| 1993 | Personnel & Guidance J |
| 2961 | Psychiatric Q |
| 4177 | J Crim Law, Crim & Police Sci |
| 4184 | Acta Americana |
| 5002 | Hospital & Community Psychiatry |
| 5041 | John Hopkins Hospital Bull |
| 5244 | Psychol Today |
| 5260 | Rev Educational Res |
| 40268 | J Laboratory & Clinical Medicine |

CLUSTER 7

| | |
|---|---|
| 687 | Archiv für Psychiatrie und Nervenkrankheiten |
| 806 | Behaviour |

| | |
|---|---|
| 851 | British J Math & Stat Psychol |
| 1995 | Personnel J |
| 4192 | Zeitschrift für Psychologie |
| 5692 | Probleme und Ergebnisse der Psychologie |
| 5766 | Pflügers Archiv |
| 6348 | Zentralblatt für die gesamte Neurologie und Psychiatrie |
| 40876 | Kybernetik |
| 42002 | Information & Control |

CLUSTER 8

| | |
|---|---|
| 177 | Advances Exper Social Psychol |
| 344 | Am J Human Genetics |
| 1543 | J Exper Social Psychol |
| 1562 | J Personality & Social Psychol |
| 4997 | Int J Addictions |
| 5076 | J Exper Analysis Behaviour |
| 5247 | Psychonomic Sci |
| 5466 | New Directions Psychol |
| 5575 | J Communication |
| 40034 | Am Naturalist |
| 40039 | Anatomical Rec |
| 40224 | Investigative Ophthalmology |
| 40614 | Biological Rev |
| 40835 | Vision Res |
| 42004 | Behaviour Res and Therapy |
| 42017 | Psychology Study Sci |

CLUSTER 9

| | |
|---|---|
| 123 | Acta Oeconomica |
| 1168 | Economics Planning |
| 1665 | Lloyds Bank Rev |
| 40299 | Kozgazdasagi Szemie |
| 40380 | Penzugyi Szemie |

CLUSTER 10

| | |
|---|---|
| 760 | Australian Economic Papers |
| 1142 | Econometrica |
| 1448 | Int Economic Rev |
| 1577 | J Am Stat Assoc |
| 1596 | J Transport Economics & Policy |
| 4809 | Annals Math Stat |
| 4928 | Economic Studies Q |
| 5301 | SIAM Rev |
| 5425 | Naval Research Logistics Q |
| 40146 | Computer J |
| 40310 | Mathematische Annalen |
| 40401 | Proc Cambridge Philosophical Soc |

CLUSTER 11

| | |
|---|---|
| 842 | Brit J Educational Psychol |
| 1203 | Educational & Psychological Measurement |
| 1206 | Educational Res |
| 1242 | Ergonomics |
| 1511 | J Applied Psychol |
| 1565 | J Projective Techniques & Personality Assessment |
| 1812 | Multivariate Behavioral Res |
| 4867 | California J Educational Res |
| 4814 | Dissertation Abstracts Int |
| 4990 | Gifted Child Q |
| 5146 | Mind (UK) |
| 5249 | Psychosomatics |
| 5292 | Scand J Psycholo |
| 40022 | Am J Epidemiology |
| 40024 | Am Medicine |
| 40047 | Annals Internal Medicine |
| 40138 | Clinical Sci |
| 40235 | J Am Medical Assoc |
| 40248 | J Clincal Investigation |
| 42010 | J Psychol Neurol |

CLUSTER 12

| | |
|---|---|
| 324 | Am Econ Rev |
| 913 | Bull Oxford Univ Inst Econ & Stat |
| 1513 | J Business |
| 1557 | J Money Credit & Banking |
| 1970 | Pakistan Development Rev |
| 2342 | Survey Current Business |
| 2488 | Yale Econ Essays |
| 3713 | J Finance |
| 4977 | Food Res Inst Stud Agriculture Econ Trade |
| 5066 | J Econ Theory |
| 5067 | J Econ & Management |
| 5194 | ODU |
| 5198 | Operations Res |
| 5250 | Public Choice |
| 5345 | Stud Income & Wealth |
| 5376 | Urban Rev |

CLUSTER 13

| | |
|---|---|
| 1554 | J Marriage & Family |
| 1564 | J Police |
| 2071 | Public Administration Rev |
| 2262 | Soc Forces |
| 2265 | Soc Sci Q |
| 2288 | Sociol Rev |
| 2293 | Sociol & Soc Res |
| 2386 | Trans-Action |
| 2400 | US News & World Report |
| 2986 | Soc Problems |
| 4175 | Human Organization |
| 4496 | Pacific Sociol Rev |
| 4816 | Atlantic Monthly |
| 4921 | Ebony |
| 5097 | J Negro Education |
| 5111 | J Sex Res |
| 5313 | Sociol Analysis |
| 5315 | Sociol Q |
| 40483 | Social Biology |

CLUSTER 14

| | |
|---|---|
| 348 | Am J Ortho |
| 1030 | Community Mental Health J |
| 1380 | Human Development |
| 1459 | Int J Psychiatry |
| 1515 | J Child Psychol & Psychi Allied Disciplines |
| 1576 | J Am Acad Child Psychol |
| 1898 | New Republic |
| 4187 | Social Work (USA) |
| 4394 | Social Case Work |
| 4953 | Exp Children |
| 5049 | J Am Public Health Assoc |
| 41019 | Am J Dis Children |

CLUSTER 15

| | |
|---|---|
| 1056 | J Advertising Res |
| 1552 | J Marketing |
| 40295 | Journalism Q |

CLUSTER 16

| | |
|---|---|
| 354 | Am J Psychol |
| 364 | Am Psychologist |
| 562 | Am Rev Psychol |
| 703 | Archives of Psychol |
| 764 | Australian J Psychol |
| 859 | Brit J Psychol |
| 959 | Canad J Psychol |
| 1346 | Harvard Ed Rev |
| 1381 | Human Factors |

CLUSTER 16 (continued)

1504 J Acoustic Soc Am
1535 J Ed Psychol
1542 J Exp Res in Person
1544 J General Psychol
1545 J Genetics
1546 J Geront
1878 Nature
2068 Psychol Bull
2074 Psychometrika
2234 Science
2967 Psychol Rev
3991 Am J Mental Def
4107 Eur J Psychonomics
4807 Am Scientist
5079 J Exp Psychol
5095 J Math Psychol
5211 Perception & Psychophysics
5212 Perception & Motor Skills
5238 Psychol Rec
5239 Psychol Rep
40012 Aerospace Medicine
40077 Bell System Tech J
40278 J Opt Soc Am
40384 Philosophische Studien

CLUSTER 17

946 Univ California Publications in Psychol
1598 J Verbal Learning & Verb Behav
1630 Language & Speech
2476 Word
4184 Q J Exp Psychol
5101 J Philosophy
5130 Language
5173 Neuropsychologia
5241 Psychologie Francais
5916 ASHA
40035 Proc Am Philosophical Soc
40058 Arch Neurol
40149 Cortex
40276 J Neurosurgery
40383 Philosophical Rev

CLUSTER 18

356 Am J Sociol
373 Am Soc Rev
374 Am Sociologist
546 Anthropological Q
553 Anthropos
697 Archives Européennes de Sociologie
1104 Current Sociol
1751 Milbank Memorial Fund Q
2031 Population Studies
2221 Rural Sociol
4909 Demography
5137 Law and Soc Rev
5187 Newsweek
5225 Population Index
5314 Sociol Inquiry
5318 Soc of Educ

CLUSTER 19

337 Am J Agricultural Econ
1151 Econ Development
1155 Econ J
1157 Econ Record
1166 Economica
1277 Federal Res Bull
1288 Foreign Affairs
1523 J Development Studies
1533 J Econ History

CLUSTER 19 (continued)

1563 J Political Econ
1609 Kyklos
1638 Law & Contemporary Problems
1707 Manchester School of Econ & Soc Studies
1947 Openbare Financien
1958 Oxford Econ Pap
2087 Q J Econ
2145 Res Econ Stud
2146 Res Income & Wealth
2263 Soc Research
2302 Southern Econ J
2348 Swedish J Econ
3864 J Econ Lit
4406 Metroeconomica
4805 Am Econ Assoc Proc
4925 Econ Bull Lat America
4927 Econ & Pol Weekly
5031 IMF Staff Papers
5054 J Asian Stud
5093 J Law & Econ
5165 Nat Banking Res

CLUSTER 20

4173 Archives de Psychologie
4191 Z f exp u angew Psych
5240 Psychologie
5242 Psychologische Forschung
5609 Diagnostica
5630 Fortschntte d Neur Psychiat
5697 Psychologische Beiträge
5698 Psychologische Rundschau
6345 Z f Tierpsychologie
40290 J Speech & Hearing Research
40499 Synthese
40576 Naturwissenschaft
40716 J Speech & Hearing Disorders

CLUSTER 21

956 Canadian J Econ & Pol Sci
1143 Economic Int.
1152 Econ Geography
1402 Int Econ & Econometrics Annual
1475 Int Soc Sci J
1621 Land Econ
2043 Princeton Studies Int Finance
3103 Pakistan Econ J
4876 Canadian J Econ
5220 Philippine Econ J
5384 Western Econ J

CLUSTER 22

355 Am J Psychotherapy
695 Archives de Psychologie
702 Arch Env Health
805 Behaviour Res & Therapy
1503 J Abnormal Psychol
1518 J Clinical Psychol
1528 J Counselling & Child Psychol
1529 J Counselling Psychol
5026 Int J Clin & Exp
5051 J App Business Analysis
5181 New Scientist
5248 Psychophysiology
6201 Psychotherapy:Theory,Res & Prac
40364 Nursing Research
41352 Progress Experimenta

CLUSTER 23

174 Advances Child Development & Behaviour

CLUSTER 23 (continued)

988 Child Development
1118 Developmental Psychol
1208 Educ Res
1460 Int J Psychol
1737 Merrill Palmer Q
2064 Psychoanalytic Stud Children
5033 Int Rev of Res into Mental Retardation
5077 J Exp Child Psychol
5234 Psychiatric Res Rep
5404 Rev Child Dev Res
5405 Reading Teacher
40279 J Pediatics
40377 Pediatics
41257 J Pharmaceutical Sci
41420 Yale J Biology Med

CLUSTER 24

306 Am Anthopologist
343 Am J Econ & Sociol
351 Am J Physical Anthropol
1096 Current Anthropol
1252 Ethnology
1454 Int J Am Linguistics
1637 Lat Am Res Rev
1691 Man
1932 Oceania
2303 SW J Anthopology
4943 Estudios de Cult Maya
40343 Natural History
40562 Bull History Medicine
40758 Phil Trans R Soc London (A)
40815 Tetrahedron Letters
41152 Evolutionary Genetics Res Rep
41163 Folia Primatologica
41221 J Dentistry Children
41376 School Sci Math

CLUSTER 25

588 Applied Stats
816 Biometrika
1165 Economic Trends
1591 J Royal Stat Soc (A)
1592 J Royal Stat Soc (B)
1948 Operational Res Q
3035 Bull Int Inst Stat Actes
5394 Yale Law J
40207 Human Biology
40210 IBM J Res & Development
40308 Mathematical Biosciences
40394 Practitioner
40421 Q J Medicine
40504 Technometrics
42001 Federation Proceedings

CLUSTER 26

133 Acta Psychiat Scand Suppl
448 Annales Médico-psychologiques
699 Arch General Psychiatry
841 Brit J Crim
852 Brit J Med Psychol
856 Brit J Prev & Soc Med
858 Brit J Psychiatry
861 Brit J Soc & Child Psychol
1455 Int J Group Psychotherapy
1458 Int J Psychoanalysis
1558 J Nervous & Mental Disease
1567 J Psychiat Res
1569 J Psychoanatic Res
1623 Lancet
1801 New Society

CLUSTER 26 (continued)

2881 Social Psychiatry
4452 Annals of Hum Gen
4878 London Psychiat Assoc J
5086 J Ment Def Rev
5254 Q J Stud Alc
5412 Comprehensive Psychiatry
5839 Psychiatric Clinica
5918 Ass Res Nerv Ment Dis
40004 Acta Medica Scandinavica
4010 British Medical J
40158 Diseases Nervous System
40165 Endocrinology
40174 Exerpta Medica Int Congress Series
40247 J Clinical Endocrinol
40408 Proc Roy Soc Med
41255 J Pharm & Pharmacol
41357 Psychology Personal

CLUSTER 27

161 Adm Sci Q
1345 Harvard Business Rev
1700 Management Science
1992 Personnel Admin
1998 Personnel Psychol
2282 Sociology
2968 Psychosomatic Medicine
5016 Int Monetary Rev
5199 Org Beh & Human Per
5295 School Life
40023 Am J Medical Sci
40056 Archives Internal Med
40196 Geriatrics
40246 J Chronic Disease
40381 Perspectives in Biology & Medicine

CLUSTER 28

1470 Industrial & Labor Relations Rev
2144 Rev Econ & Stats
2169 Rev Inst Int Stat
2268 Soc Security Bull
4959 Fed Res Bank of St. Louis Rev
5080 J Fin & Quantative Anal
5170 Nat Tax J
5576 J Regional Science
40141 Communications ACM

CLUSTER 29

784 Banca del Curaçao Q Rev
848 Brit J Ind Rel
1177 Economist
1463 Int Lab Rev
1548 J Ind Econ
1782 Monthly Bull Stats
1893 Moorgate & Wall St
1853 Nat Inst Econ Rev
5344 Studies in Family Planning
6429 Bull Econ Res
40286 J des Savants

CLUSTER 30

363 Am pol Sci Rev
469 Ann Assoc Amer Geog
1054 Comparative Stud in Soc & Hist
1154 Econ History Rev
1527 J Conflict Resol
1585 J Market Res Society
2025 Pol Studies
2081 Public Opinion Q
2996 World Politics

CLUSTER 30 (contined)

4998 Daedalus
4955 Harvard Law Review
5144 Independent J Pol Sci
5183 NY Times Mag
5385 Western Political Q

Table 12

Possible labels for 30 clusters from the ranked sample

| | |
|---|---|
| 1. Geography, etc | 16. Psychology |
| 2. Foreign | 17. Language and Speech |
| 3. Behavioural Science | 18. Sociology and Anthropology |
| 4. Education, etc | 19. Economics |
| 5. Brain Research | 20. German Psychology |
| 6. Social Psychiatry | 21. Economics |
| 7. German | 22. Clinical Psychology |
| 8. Applied Psychology | 23. Experimental Child Psychology |
| 9. Economics | 24. Anthropology |
| 10. Mathematical Economics | 25. Statistics |
| 11. Education and Medicine | 26. Social Medicine in Britain |
| 12. Business Economics | 27. Personnel Psychology, etc |
| 13. Sociology | 28. Economics and Labour |
| 14. Children | 29. Labour Relations |
| 15. Marketing | 30. Economic History |

Figure 8 shows the pattern in which the clusters merged and the main occasions that the merging departed from hierarchical. The general pattern is difficult to see from the figure, but there are two main branches to the tree and a number of rather isolated clusters. The main branches are as follows.

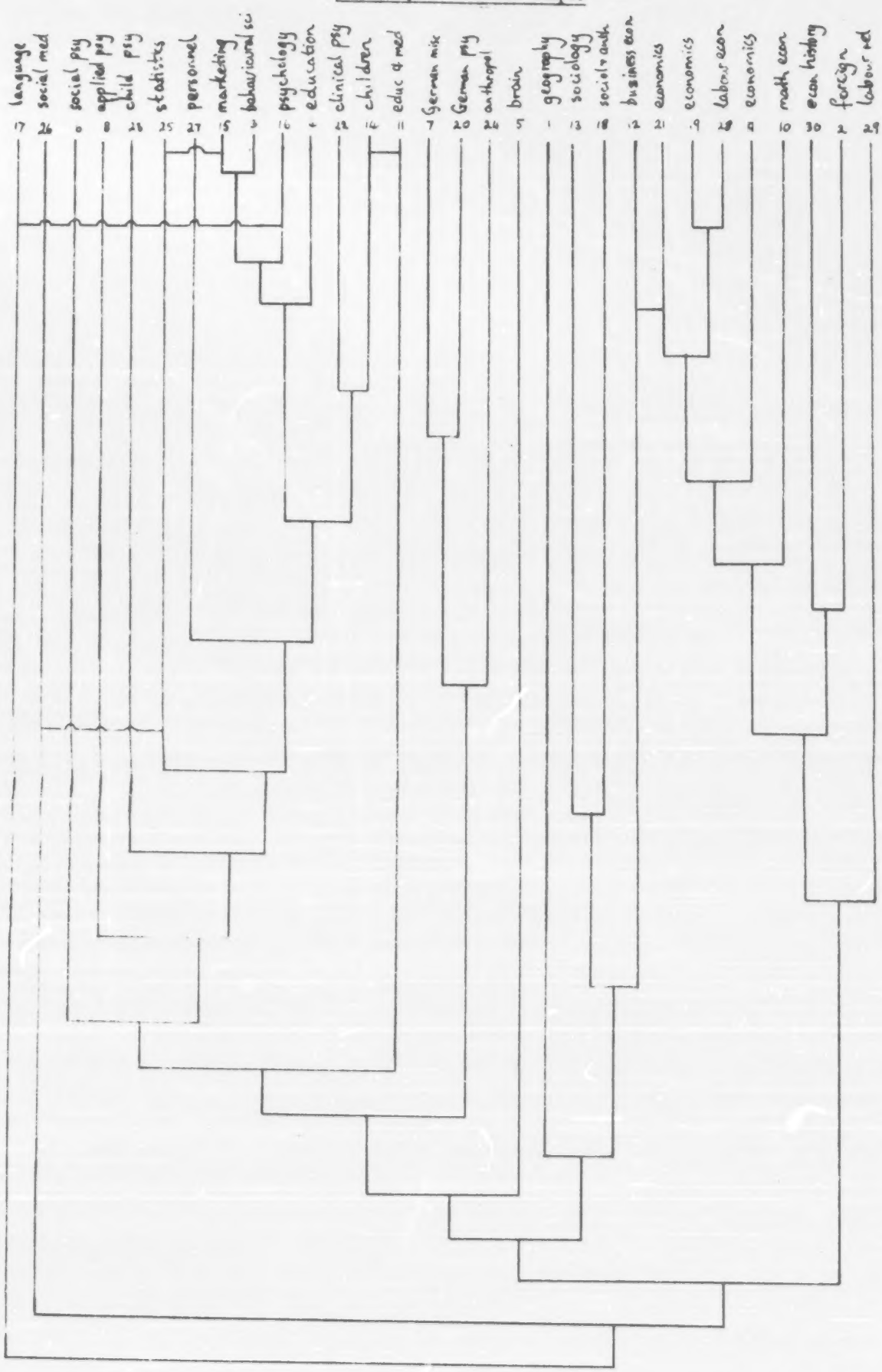
(i) Clusters 6, 8, 23, 25, 27, 15, 3, 16, 4, 22 and part of 14. These are in fields of psychology and its various applications and related techniques.

(ii) Clusters 21 (part), 19, 28, 9, 10, 30, 2 and 29. These are all in economic fields.

Of the isolated clusters, numbers 1 (geography), 5 (brain research), 13 (sociology), 19 (sociology and anthropology) and 24 (anthropology) are particularly distinct. Two other very distinct clusters cover the German language journals.

Figure 8

Hierarchy from Ranked Sample



Between 30 and 10 clusters there were 111 non-hierarchical jumps, that is, at the 10 cluster level, about 11 per cent of the journals were not in the cluster that would have been expected from the hierarchy. The following table lists the original clusters and the number of jumps out of each.

Table 13

Ranked sample - number of non-hierarchical jumps

| Original cluster number | Number of jumps | Original cluster number | Number of jumps |
|---|---|---|---|
| 1 | 2 | 16 | 14 |
| 2 | 0 | 17 | 1 |
| 3 | 17 | 18 | 3 |
| 4 | 9 | 19 | 5 |
| 5 | 2 | 20 | 2 |
| 6 | 1 | 21 | 9 |
| 7 | 2 | 22 | 3 |
| 8 | 0 | 23 | 2 |
| 9 | 1 | 24 | 1 |
| 10 | 0 | 25 | 0 |
| 11 | 2 | 26 | 4 |
| 12 | 2 | 27 | 1 |
| 13 | 0 | 28 | 7 |
| 14 | 18 | 29 | 0 |
| 15 | 4 | 30 | 1 |

The number of non-hierarchical jumps is an important indicator of how homogeneous are the individual clusters. It also indicates what would be the effect of using a hierarchical clustering method. At first glance 11 per cent does not seem a large number of jumps but there were two places where a cluster splits almost in half. These can be seen on Figure 8.

In the first instance cluster 14, on children, is divided into two parts. The first merged into cluster 11, on education and medicine; the second continued separately and later merged with cluster 22, on clinical psychology.

The other cluster to split into two parts was cluster 16, on psychology. A large part merged into cluster 17, on language and speech,

while the remainder continued and later joined up with clusters on education, behavioural science and marketing.

## 6.2 Citations from the random list of journals

The journals cited by the random list of sources were clustered and analysed in the same way as the journals from the ranked list of sources. Table 14 gives the entire clustering at the 30-cluster level. Even more than with the ranked sample several of the clusters were difficult to label. Table 15 is an attempt to label the clusters, but three are too mixed to identify.

Figure 9 shows the pattern in which the clusters merged and the main occasions that the merging departed from hierarchical. The principal branches of this hierarchy are as follows.

(i) Clusters 15, 24, 2 and 29. These are all in psychiatry or medicine. Except for a few journals in social psychology these form a tight cluster.

(ii) Clusters 1, 14, 8, 25, 16 and 7. These are all in psychology.

(iii) Clusters 18, 30, 27 and 9. Except for two journals in cluster 30 these are all on language and linguistics.

Apart from these main branches, several small groups of clusters are quite distinct. They are numbers 21, 22 and 13 on sociology, education, etc; 6 and 20 on economics; and 30, 10 and 26 which are a mixed group of geography, market research and operations research.

Between 30 and 10 clusters there were 53 non-hierarchical jumps, that is, at the 10-cluster level, about 7 per cent of the journals were not in the cluster that would have been expected from the hierarchy. The following table lists the original clusters and the number of jumps out of each.

Table 14

30 clusters from random sample

CLUSTER 1

| | |
|---|---|
| 815 | Biometrics |
| 841 | Brit J Criminology |
| 859 | Brit J Psychol |
| 1341 | Group Psychotherapy |
| 1937 | Occupational Psychol |
| 40315 | Medical Press |

CLUSTER 2

| | |
|---|---|
| 133 | Acta Psychiat Scand, Supp |
| 352 | Am J Psychiatry |
| 699 | Arch General Psychiatry |
| 702 | Arch Environmental Health |
| 858 | Brit J Psychiatry |
| 959 | Canadian J Psychol |
| 1503 | J Abnormal Psychol |
| 1558 | J Nervous & Mental Disease |
| 1559 | J Neurology, Neurosurgery and Psychiatry |
| 1567 | J Psychiatric Res |
| 2968 | Psychosomatic Medicine |
| 40026 | Am J Obstetrics & Gynecology |
| 40056 | Arch Int Medicine |
| 40279 | J Pediatrics |
| 40280 | J Physiology |
| 40290 | J Hearing & Speech Res |
| 40324 | Metabolism |
| 40352 | Neurology |
| 40408 | Proc Roy Soc Medicine |
| 40410 | Proc Soc Exp Biology & Medicine |
| 40750 | Obstetrics & Gynecology |

CLUSTER 3

| | |
|---|---|
| 468 | Ann Assoc Amer Geog |
| 956 | Canad J Econ & Pol Sci |
| 1591 | J Roy Statistical Soc (A) |
| 1621 | Land Economics |
| 2129 | Reg Sci Assoc Papers |
| 2970 | J Am Inst Planners |
| 4872 | Canadian Geographer |
| 5198 | Operations Research |
| 5576 | J Reg Sci |
| 40271 | J Math Analysis & Applications |

CLUSTER 4

| | |
|---|---|
| 468 | Annals Am Ac Pol & Soc Sci |
| 1008 | College & Res Libraries |
| 2753 | Library Journal |
| 4421 | Business Week |
| 5017 | Information Bulletin |

CLUSTER 5

| | |
|---|---|
| 4189 | Voprosy Ekonomiki |
| 6137 | Nedepya |
| 6301 | Uchenye Zapiski Uralskii Gos Univ |
| 6313 | Vestnik Moskovskogo Universiteta Seriya Geografiya |
| 40576 | Naturwissenschaft |
| 40619 | Botanicheskii Zhurnal |
| 40746 | Novoe v Biologii i Meditsine |

CLUSTER 6

| | |
|---|---|
| 1142 | Econometrica |
| 1155 | Economic Journal |
| 1157 | Economic Record |
| 1166 | Economica |
| 1448 | Int Economic Rev |

CLUSTER 6 (continued)

| | |
|---|---|
| 1563 | J Polit Econ |
| 1958 | Oxford Economic Papers |
| 2087 | Q J Economics |
| 2145 | Rev Econ Studies |
| 2488 | Yale Economic Essays |
| 3713 | J Finance |
| 5080 | J Financial & Quantitative Anal |

CLUSTER 7

| | |
|---|---|
| 161 | Admin Sci Q |
| 803 | Behavioral Sci |
| 1345 | Harvard Bus Rev |
| 1413 | Industrial Relations |
| 1527 | J Conflict Resolution |
| 1998 | Personnel Psychol |
| 2295 | Sociometry |
| 4175 | Human Organization |
| 4809 | Ann Math Stats |
| 40023 | Am J Medical Sciences |
| 40030 | Circulation |
| 40235 | J Am Medical Ass |
| 40248 | J Clinical Invest |
| 40641 | Clincial Chem Acta |

CLUSTER 8

| | |
|---|---|
| 354 | Am J Psychol |
| 703 | Archives of Psychol |
| 988 | Child Devel |
| 1545 | J Genet Psychol |
| 1546 | J Gerontol |
| 1737 | Merrill Palmer Quarterly |
| 4191 | Z f exp u angew Psychologie |
| 4806 | Am J Disturbed Ch |
| 5077 | J Exp Child Psychol |
| 5242 | Psych Forsch |
| 5292 | Scand J Psychol |
| 40077 | Bell Syst Tech J |
| 40196 | Geriatrics |

CLUSTER 9

| | |
|---|---|
| 219 | African Studies |
| 1454 | Int J Am Linguistics |
| 1600 | J West Afr Languages |
| 4907 | Current Topics Linguistics |

CLUSTER 10

| | |
|---|---|
| 337 | Am J Agric Econ |
| 1553 | J Marketing Res |
| 1609 | Kyklos |
| 4193 | Z für Wirtschafts und Sozialwiss |

CLUSTER 11

| | |
|---|---|
| 362 | Am Philosophical Q |
| 2444 | Voprosi Filosofi |
| 5180 | New Politics |
| 5380 | Voprosy Psikhologii |

CLUSTER 12

| | |
|---|---|
| 944 | Cahiers Internationaux de Sociologie |
| 1296 | Francais Moderne |
| 2171 | Revue de Linguistique Romane |

CLUSTER 13

2008 Philos of Science
5146 Mind (UK)
5577 Mens en Maatschaapij
40499 Synthese

CLUSTER 14

364 American Psychol
1203 Educ & Psych Measurement
1383 Human Relations
1511 J Appl Psychol
1529 J Counsel Psychol
1542 J Exp Res in Pers
1543 J Exp Soc Psychol
1561 J Personality
1562 J Personality & Soc Psychol
1568 J Psychology
1575 J Soc Psychol
2065 Psychological Bull
2237 Scientific American
2967 Psychol Rev
3396 J Abnorm & Soc Psychol
4807 American Scientist
5059 J Comp & Phys Psychol
5079 J Exp Psychol
5212 Perceptual & Motor skills
5239 Psychol Reports
5247 Psychonomic Science

CLUSTER 15

856 Brit J Prev & Soc Med
1515 J Ch Psychol & Psychi & Allied Disciplines
1623 Lancet
1878 Nature
2988 Soc Sci & Medicine
3991 Am J Mental Def
40045 Annals of Allergy
40101 Brit Medical J
40114 Canad Medical Assoc J
40247 J Clinical Endocrinology
40377 Pediatrics
40584 Acta Paediatrica
40716 J Speech & Hearing Disorder
40757 Pharmacological Rev

CLUSTER 16

912 Bull Mental Clinic
1346 Harvard Educ Rev
1554 J Marriage & Family
1732 Mental Hygiene
2061 Psychiatric Q
2065 Psychoanalytic Rev
2069 Psychol Issues
2400 US News & World Rep
40158 Disorders Nervous System
40581 Acta Anatomica
40759 Physiol Rev

CLUSTER 17

849 Brit J Industrial Relations
1151 Economics Development & Cultural Change
1513 Journal of Business
1533 J Econ History
1576 J Transport Econ & Pol
1665 Lloyds Bank Review
1794 Monthly Labor Review
2260 Social Econ Admin
5016 Industrial Management Rev

CLUSTER 18

1294 Foundation of Language
1630 Language & Speech
2476 Word
4184 Q J Exp Psychol
5101 J Philosophy
5129 Languages
5130 Language
6097 Linguistique Générale

CLUSTER 19

1133 Droit Social
5634 Giusticia Civile
6023 Giur Compl Can Civ

CLUSTER 20

324 Am Econ Rev
2144 Rev Econ & Stats
2302 Southern Econ J
4811 Arch J

CLUSTER 21

218 African Social Research
356 Am J Sociology
363 Am Pol Sci Rev
373 Am Soc Rev
697 Arch Européennes de Sociologie
864 Brit J Sociology
1753 Minerva
2006 Philosophy
2262 Social Forces
2263 Social Research
2292 Sociology
2634 Africa
2881 Proc Aristotelian Soc
5093 J Law and Economics
5181 New Scientist
5574 Inquiry

CLUSTER 22

842 Brit J Education Psychol
1536 J Education Res
1556 J Modern African Studies
2221 Rural Sociology
40434 Research Q

CLUSTER 23

1237 English Studies
1295 Français dans le Monde
1663 Listener
40395 Praxis

CLUSTER 24

| | |
|---|---|
| 131 | Acta Psychiatrica Scand |
| 4107 | European J Psychonomics |
| 5407 | Psychopharmacologia |
| 40094 | Brain |
| 40312 | Med J of Australia |

CLUSTER 25

| | |
|---|---|
| 805 | Behaviour Research & Therapy |
| 1518 | J Clinical Psychol |
| 1528 | J Consulting & Clinical Psychol |
| 1535 | J Educational Psychol |
| 1565 | J Projective Techniques & Personality Assessment |
| 2062 | Psychiatry |
| 2074 | Psychometrika |
| 4187 | Social Work (USA) |
| 4394 | Social Casework |
| 5002 | Hospital & Community Psychiatry |
| 5051 | J Applied Behavior Analysis |

CLUSTER 26

| | |
|---|---|
| 1552 | J of Marketing |
| 1577 | J Am Stat Assoc |
| 1585 | J Market Research Soc |
| 1700 | Management Science |
| 1948 | Operational Research Quarterly |
| 2081 | Public Opinion Quarterly |
| 5425 | Naval Research Logistics Quarterly |
| 40146 | Computer Journal |
| 40479 | SIAM Journal of Applied Maths |

CLUSTER 27

| | |
|---|---|
| 1651 | Lingua |
| 4856 | Bull Soc Linguistique de Paris |

CLUSTER 28

| | |
|---|---|
| 4908 | Daedalus |
| 5183 | New York Times Magazine |
| 5187 | Newsweek |
| 5291 | Saturday Review |

CLUSTER 29

| | |
|---|---|
| 348 | Am J Orthopsychiatry |
| 357 | Am J Public Health |
| 861 | Brit J Social Clinical Psychol |
| 1544 | J General Psychol |
| 1573 | J Social Issues |
| 1691 | Man |
| 1751 | Milbank Memorial Fund Q |
| 1901 | New Society |
| 1993 | Personnel & Guidance J |
| 2234 | Science |
| 5249 | Psychosomatics |
| 40049 | Annals NY Acad Science |
| 40246 | J Chronic Disease |
| 40355 | New England J Medicine |

CLUSTER 30

| | |
|---|---|
| 306 | Am Anthropologist |
| 1657 | Linguistics |
| 5830 | Revue Philosophique de la France et de l'Etranger |

Table 15

Possible labels of 30 clusters from the random sample

| | |
|---|---|
| 1. Psychology | 16. Social Psychiatry |
| 2. Psychiatry & Medicine | 17. Business Economics |
| 3. Geography | 18. Language |
| 4. Information & Libraries | 19. Foreign |
| 5. Russian | 20. Economics |
| 6. Economics | 21. Sociology |
| 7. Behavioural Science | 22. Education |
| 8. Child Psychology | 23. Mixed |
| 9. African Languages | 24 Psychiatry |
| 10. Mixed | 25. Psychiatry and Clinical Psychology |
| 11. Philosopy | 26. Operational Research |
| 12. French | 27. Linguistics |
| 13. Epistemology | 28. News Magazines |
| 14. Experimental Psychology | 29. Social Psychology |
| 15. Medicine | 30. Mixed |

## Figure 9

## Hierarchy from random sample

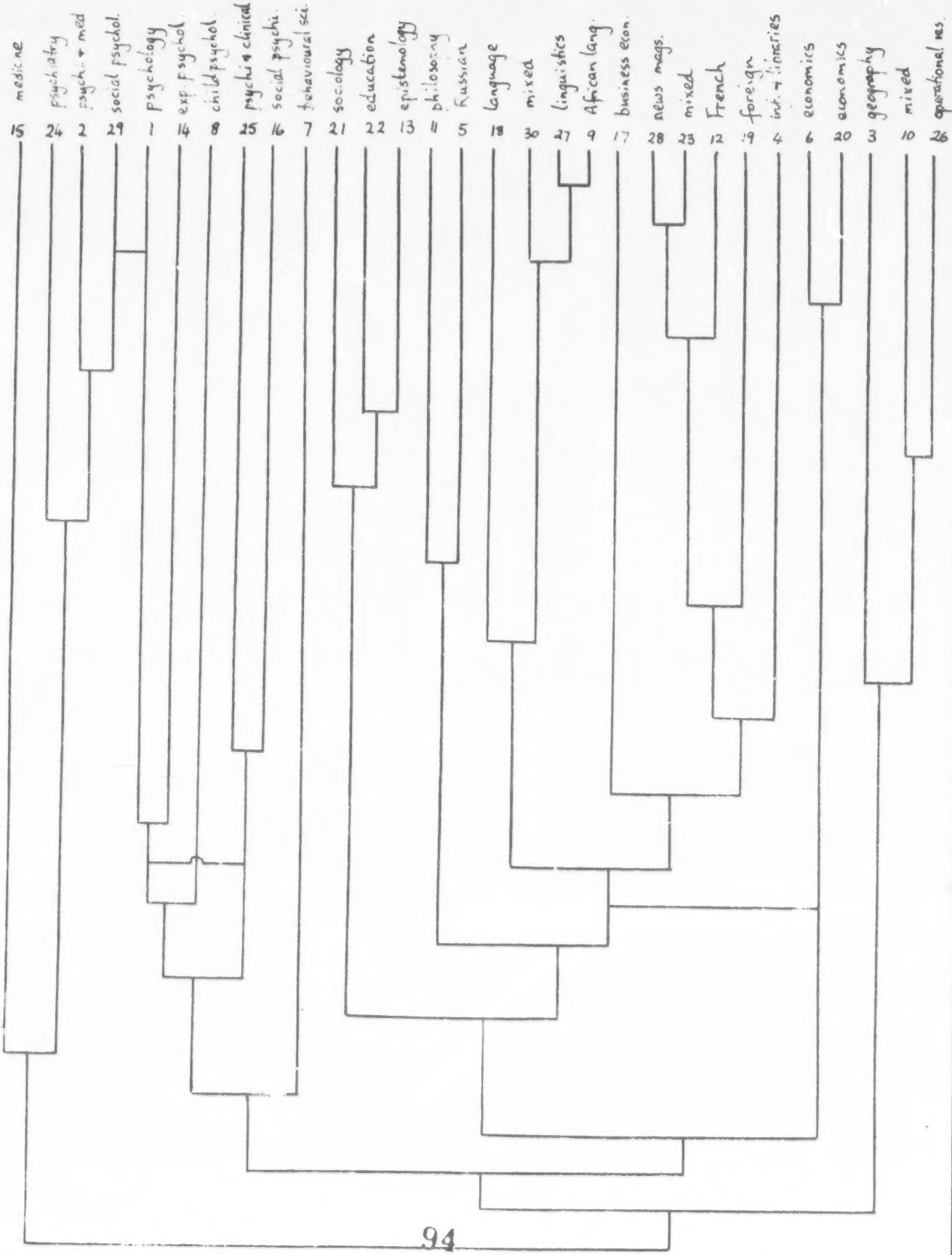


Table 16

Random sample - number of non-hierarchical jumps

| Original cluster number | Number of jumps | Original cluster number | Number of jumps |
|---|---|---|---|
| 1 | 4 | 16 | 0 |
| 2 | 0 | 17 | 2 |
| 3 | 0 | 18 | 0 |
| 4 | 1 | 19 | 0 |
| 5 | 1 | 20 | 0 |
| 6 | 0 | 21 | 0 |
| 7 | 0 | 22 | 5 |
| 8 | 4 | 23 | 1 |
| 9 | 2 | 24 | 1 |
| 10 | 5 | 25 | 6 |
| 11 | 1 | 26 | 3 |
| 12 | 0 | 27 | 0 |
| 13 | 1 | 28 | 3 |
| 14 | 1 | 29 | 11 |
| 15 | 0 | 30 | 1 |

The most important instance of the clustering not following the hierarchy involves cluster 29. Originally this is a mixture of sociology medicine and applied psychology. It splits into many pieces. The two largest form parts of the main branches of the tree which cover medical psychiatry and psychology respectively, but other journals are widely scattered.

## 6.3 Comparison between the clusters derived from the ranked and random samples

The purpose of using two separate sets of data with sources selected by different criteria was to see how robust the results were. In particular it is interesting to compare the following for the two samples.

(i) The overlap between the journals cited.

(ii) The subject distributions of the journals cited.

(iii) For those journals that are in both samples, the clusters that they fall into.

The third of these is particularly important, indicating the degree of robustness. Table 10 on page 69 shows the overlap of journals between the two samples. The first point of interest is that although the numbers of source journals were almost exactly the same, 495 journals were cited by at least two of the journals in the ranked list, but only 254 were cited by at least two of the journals in the random list. This may be due to the ranked sample containing more citations because the core journals in the ranked list tend to have more issues per year and more pages per issue. Even so the difference is spectacular.

On the other hand, of the journals in the random sample, no less than 71 per cent were also in the ranked sample, which seems to be a very substantial overlap.

Tables 17 and 18 give an approximate breakdown of the journals in the two samples by subject area. Since many journals cover more than one subject area the totals exceed the total number of journals. In this table the subject codings for the source journals are reliable, but many of the cited journals have not been examined and have been coded from their titles.

Table 17

Subjects of journals in the ranked sample

| | Source journals | | Cited journals | |
|---|---|---|---|---|
| | Number | per cent | Number | per cent |
| Anthropology | 3 | 6 | 17 | 3 |
| Criminology | 1 | 2 | 3 | 1 |
| Economics | 13 | 27 | 75 | 15 |
| Education | 1 | 2 | 18 | 4 |
| Environmental Planning | 0 | 0 | 1 | 0.2 |
| Geography | 0 | 0 | 13 | 3 |
| Librarianship | 0 | 0 | 0 | 0 |
| Linguistics | 1 | 2 | 5 | 1 |
| Management | 2 | 4 | 27 | 6 |
| Politics | 1 | 2 | 16 | 3 |
| Psychology | 22 | 46 | 131 | 26 |
| Social Welfare | 1 | 2 | 15 | 3 |
| Sociology | 5 | 10 | 39 | 8 |
| Statistics | 2 | 4 | 16 | 3 |
| Others | 2 | 4 | 130 | 26 |
| TOTAL | 48 | 100 | 495 | 100 |

Table 18

Subjects of journals in the random sample

| | Source journals | | Cited journals | |
|---|---|---|---|---|
| | Number | per cent | Number | per cent |
| Anthropology | 0 | 0 | 1 | 0.4 |
| Criminology | 1 | 2 | 1 | 0.4 |
| Economics | 3 | 6 | 25 | 10 |
| Education | 5 | 11 | 5 | 2 |
| Environmental Planning | 3 | 6 | 2 | 1 |
| Geography | 3 | 6 | 9 | 4 |
| Librarianship | 2 | 4 | 3 | 1 |
| Linguistics | 6 | 13 | 16 | 6 |
| Management | 1 | 2 | 16 | 6 |
| Politics | 8 | 17 | 18 | 7 |
| Psychology | 8 | 17 | 66 | 25 |
| Social Welfare | 1 | 2 | 7 | 3 |
| Sociology | 8 | 17 | 19 | 7 |
| Statistics | 1 | 2 | 10 | 4 |
| Others | 3 | 6 | 73 | 29 |
| TOTAL | 49 | 100 | 254 | 100 |

These two tables show that the source journals in the random sample were more widely spread across subject areas than those in the ranked sample. Over 70 per cent of the sources in the ranked sample were in psychology. The cited journals were more widely spread than the sources. A substantial number of medical and other scientific journals were cited, usually by psychology or psychiatry sources.

The difference in the number of economics journals, both sources and cited, in the two samples accounts for one of the biggest differences between the two sets of clusters; the ranked sample has a major branch of several economics clusters which is largely missing in the random set of clusters. The wider subject spread also accounts for several of the random clusters being less homogeneous than the ranked clusters.

Table 19 is one of the key tables. For the 184 journals which are cited by sources from both the ranked and random lists it compares the clustering at the 20-cluster level. For each cluster it shows how many journals lie in each cluster of the other sample. The cluster level 20 was chosen for this table so as to eliminate any bias from the starting up procedure used by the clustering program.

The table shows a pattern in which each cluster from one sample tends to be divided among a small number of cluster from the other. For example, the 16 journals from clusters 4 of the ranked sample occur in four clusters from the random sample. Two conclusions can be drawn from this table.

(i) The structures identified by the two sets of data have a large number of common features. This is shown by the small number of non-zero entries in this table.

(ii) The structure is not sufficiently clear cut to be confident in assigning each individual journal to a specific cluster. The two sets of data show that many journals could reasonably lie in one of two or more clusters. Without much more data the allocation is slightly arbitrary.

Table 19

Comparison of ranked and random allocations

| Random \ Ranked | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | 2 | 2 | | 1 | | | | | 1 | | | | | | | | | | 6 |
| 2 | | | 5 | 8 | 7 | | | 2 | | 3 | | | | | | | 1 | 1 | 1 | 2 | 30 |
| 3 | | | 1 | | | | 1 | | 3 | | | 2 | | 1 | 1 | | | | | | 9 |
| 4 | | | | | | | | | 1 | | | | | | 1 | | | | | | 2 |
| 5 | | | | | | | | | | | | | | | | | | | | | 0 |
| 6 | | | | | | | | | 11 | | | 3 | | | | | | | | | 14 |
| 7 | | | 2 | | | | 1 | | 1 | 5 | 2 | | 1 | | 1 | | | | | | 13 |
| 8 | | 1 | 5 | | | | 2 | | | 1 | 1 | | | | | | | | 3 | | 12 |
| 9 | | | | | 1 | | | | | | | | | | | | | | | 2 | 2 |
| 10 | | | | | | | | | 5 | 1 | | | | 1 | 2 | 1 | | | | | 10 |
| 11 | | | | | | | | | | | | | | | | | | | | | 0 |
| 12 | | | | | | | | | | | | | | | | | | 1 | | | 1 |
| 13 | 1 | | 3 | | | | | | 2 | 1 | | | 3 | | 2 | | | 3 | | | 15 |
| 14 | 1 | | 12 | | 1 | | | 4 | | | 3 | | | | | | 1 | | | | 22 |
| 15 | | | 2 | 5 | | | 2 | 1 | | | 1 | | | | | | | | 1 | | 12 |
| 16 | | | 3 | 1 | 1 | 4 | | | | | | | 2 | | 1 | | | | | | 12 |
| 17 | | | | | | | | | 5 | 1 | | 1 | | | | 1 | | | | | 8 |
| 18 | | | | | | | | | | | | | | | | | 5 | | | | 5 |
| 19 | | | | | | | | | | | | | | | | | | | | | 0 |
| 20 | | | 8 | | | 2 | | | | | 1 | | | | | | | | | | 11 |
| TOTAL | 2 | 0 | 43 | 16 | 9 | 7 | 6 | 7 | 28 | 12 | 9 | 6 | 6 | 2 | 8 | 2 | 7 | 5 | 5 | 4 | 184 |

## 7.0 DISCUSSION OF RESULTS

### 7.1 Reliability and stability of clusters

If the results of clustering journal titles are to be used in the design of secondary services it is essential to assess the reliability of the clusters. There appear to be three main factors which affect the clusters:

(i) date of source journals
(ii) selection of source jounals
(iii) clustering method.

#### 7.1.1 Date of source journals

The stability of clusters over time was studied by collecting citations from the same source journals from a number of different years and comparing the clusterings which are obtained using each year's data separately. It is obviously not feasible to collect citations from a large number of journals in a large number of years, but data was collected from criminology journals for 1950, 1960 and 1970.

Narin and Carpenter (1972) for scientific journals and Xhignesse and Osgood (1967) for psychology journals have both found a considerable amount of stability in citation structure over time, although it is clearly affected by the emergence of new journals and the disappearance of old ones. The criminology data did not bear this out. Although the three runs showed basic similarities the differences between 1950, 1960 and 1970 were large. This probably results from the diffuse nature of criminology and its large overlap with other subject areas, in contrast with the more clear cut and stable areas covered by the earlier studies.

#### 7.1.2 Choice of sources

The choice of source journals for the citation collection clearly affects the frequency of citation of cited journals and thus the clusters obtained. This effect is obviously greater if only a few source journals are used. The number of citations and the number of sources in the main citation file were large by the usual standards of citation studies and might have been expected to be sufficient to give reliable clusters, but this can be questioned.

The distribution of the number of times that each journal is cited is highly skewed. Although a few journals are heavily cited, many more

are cited rarely. For example, in the ranked sample the most cited journals were the Journal of Experimental Psychology (842 citations) and the American Economic Review (584 citations). On the other hand, even after excluding those journals cited by only one source, more than half the remainder were cited less that 10 times. A different set of sources or different data from the same sources could make a considerable difference.

Different sets of source journals may produce very different clusterings, but only subjective judgement can be used to select one as superior, just as subjective judgement enters into the selection of source journals. It is reassuring that the clustering using the randomly chosen source journals gave results essentially similar to those obtained with the source journals in the ranked list from the pilot study. However this is not proof that the clusters using either list are in some way valid nor that one is more valid than the other.

Of the two methods used to select source journals, the ranked list of social science journals known to be heavily cited was expected to be more satifactory than the random set of titles. However, the ranked sample had one disadvantage; it tended to over-emphasize subject areas where authors cite heavily and to concentrate on the major subjects to the disadvantage of small areas. In particular, large numbers of psychology and economics journals were included.

The random list of journals was spread over a wider range of subjects. Random sampling poses a number of problems, particularly in choosing a sampling frame, and there is always the possibility that a subject area is completely unrepresented, as happened to anthropology. A possible test of reliability of the chosen set of sources might be to divide the list of journals into two sections, trying as far as possible to maintain the distribution over discipline, language, etc. If these two sets of source journals produced results similar to each other and to those obtained using the combined set, then a great deal of confidence could be placed in the reliability of the results in so far as selection of source journals was concerned. Another interesting possibility would be to keep the same list of sources but partition the citations.

This study was restricted to citations from journals to other serials, yet social science literature is much wider and in some subject areas less than half of all citations are to serials. Citations from

monographs and to other types of publication are discussed in two other DISISS Reports A3 and B7. They are important but there is not simple way of handling them.

7.3.1 Clustering methods

The effect of particular clustering methods on the clusters obtained from a given set of data has been discussed in Section 2 and Section 4.2. As mentioned before, Narin and Carpenter (1972) made the assumption that clusters which were found by several methods were reliable clusters. The results using the pilot study data with the SCICON methed and single and complete linkage suggest that the main clusters identified by the SCICON method are also detected by the other approaches. It is also necessary to consider a sequence of results obtained by the SCICON method for different numbers of clusters, to get an idea of the stability of clusters within the process. A cluster at one level which distributes itself between several different clusters when the number of clusters is reduced is clearly not stable even within the data collected and the clustering process used. It is worth stating again that the SCICON criterion tends not to produce clusters of very different sizes at any one level. This means that the effect of outlying points may be large and omissions of obvious outliers from the clustering runs should be considered.

Two factors dominated the results:

(i) a non-hierarchical clustering method was used;
(ii) social science subject areas overlap extensively.

The results show conclusively that hierarchical methods are not suitable for social science information. Although it is always possible to impose a hierarchical structure on any data this would be unwise. The manner in which clusters tend to split and reform implies that to impose a hierarchy would be to invent a structure that is not in the data.

Again the overlapping of subject areas means that the determination to find distinct clusters is probably unwise. Some journals naturally fall into more than one cluster and others, such as New Society, do not

naturally fit into any cluster. This study had hoped to look at overlapping clusters which are a natural extension of the SCICON algorithm. However this extension would have required vast amounts of computer time and was not carried out.

## 7.2 Evaluation of clusters

Evaluation of clusters is possible only within the context of the use to which the results are to be put. Hence clusters of articles used in the design of an information retrieval system can be evaluated only by testing them with user requests, either in a real system over a period of time by monitoring user reaction, or by simulating the results using selected requests for which recall and relevance ratios (or some similar concepts) can be measured.

It is not immediately obvious how clusters of journals intended as a basis for secondary service coverage can be evaluated. Comparison with conventional subject classification would be interesting, but if there are substantial differences a decision (which implies an evaluation) is necessary as to which breakdown is more suitable for selecting journals to cover in a particular secondary service. An estimate of the proportion of citations from journals within the cluster to journals outside that cluster would be useful.

### 7.2.1 Representation of clusters

Associated with the problem of evaluating clusters is that of representing them in such a way that the maximum amount of useful information is extracted from the results, bearing in mind the aims of the clustering. Long lists of journals which form the basic output from the computer program are cumbersome and impossible to assimilate without some condensing and subsidiary analysis. Diagrammatic representation is difficult, as the Euclidean space in which the clusters are embedded is of more than two dimensions. However, it is usually easier to get information from diagrams than from tables of figures.

The diagrams in this report are attempts to represent very complicated data. Despite the manual effort put into them, they are not always successful in attaining the twin aims of being:

(i) comprehensible;

(ii) an unbiased summary of the results.

Representing overlapping clusters would be much more difficult.

## 7.3 The use of cluster analysis as a practical tool

Cluster analysis of citations is probably not a practical aid to designing secondary services in the social sciences. First there are physical difficulties. This study collected over 50,000 citations, but still more citations would have been useful. Perhaps to only 100 journals were there sufficient citations to be confident of having achieved a reasonable sample. The cost and time required to collect, check and input this volume of data are considerable. The data is likely to be months or, more likely, years old before it is consolidated and will have an appreciable error rate caused by the near impossibility of reconciling different abbreviations of the same title.

The second problem lies in interpreting the results. Large volumes of data with a complex structure produce complicated results. This is not unique to cluster analysis. It is a common problem of multivariate statistics.

The final problem comes from overlap among social sciences. It is clear that, although the general pattern that emerges is independent of the precise data and method chosen, the specific results are more sensitive and are in no way absolute.

However cluster analysis does seem to have a role in areas where statistical retrieval is more important than specific results. Automatic retrieval systems are the obvious example. Whether the results obtained are sufficiently better than sympathetic hand analysis is another question which will depend on the specific application.

## APPENDIX I

Criterion used in the SCICON Algorithm for optimizing the division of the points into clusters.

The criterion minimized is the root mean square deviation of points from the centres of the clusters to which they have been allocated.

$$C = \left( \frac{1}{(N-M)} \sum_{i=1}^{N} d_i^2 \right)^{\frac{1}{2}}$$

where N is the total number of points,

M is the number of clusters into which the points are divided.

$d_i$ is the distance from point i to the centre of the cluster to which it is allocated.

Because the average distance of a point from its cluster centre tends to be greater if the number of clusters is reduced, values of the criterion for clusterings into different numbers of clusters are not directly comparable. However, consideration of the change in criterion as the number of clusters is reduced will reveal any very distinct optimum number of clusters.

At any stage of the algorithm M is fixed and since $C^2$ increases monotonically with C, it is sufficient to minimize the sum of squares of deviations of points from their cluster centres.

$$C' = \sum_{i=1}^{N} d_i^2$$

The change in this expression when a point is moved from one cluster to another is particularly simple. This fact accounts for the efficiency of the algorithm.

Using Euclidean distance, $d_i^2 = \sum_{j=1}^{J} d_{ij}^2$ where J is the number of dimensions, and $d_{ij}$ is the distance of point i from the cluster centre measured along the jth axis.

Hence $$C' = \sum_{i=1}^{N}\left(\sum_{j=1}^{J} d_{ij}^2\right) = \sum_{j=1}^{J}\left(\sum_{i=1}^{N} d^2{}_{ij}\right)$$

We can therefore conveniently consider the change in a single dimension.

Given a cluster of n elements, the cluster centre is at

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Adding a further element $x_s$ to this cluster moves the centre to

$$\bar{x}' = \frac{1}{n+1}\left\{\sum_{i=1}^{n} x_i + x_s\right\}$$

$$= \frac{n\bar{x} + x_s}{(n+1)}$$

The increase in the contribution of this cluster to C' is

$$\Delta C_+ = \sum_{i=1}^{n}\left\{(x_i - \bar{x}')^2 - (x_i - \bar{x})^2\right\} + (x_s - \bar{x}')^2$$

Now $$\sum_{i=1}^{n}\left\{(x_i - \bar{x}')^2 - (x_i - \bar{x})^2\right\} = n\bar{x}'^2 - 2\bar{x}'\sum_{i=1}^{n} x_i - n\bar{x}^2 + 2\bar{x}\sum_{i=1}^{n} x_i$$

$$= n(\bar{x}'^2 - 2\bar{x}'\bar{x} + \bar{x}^2) = n(\bar{x}' - \bar{x})^2$$

$$= \frac{n}{(n+1)^2}(n\bar{x} + x_s - (n+1)\bar{x})^2$$

$$= \frac{n(x_s - \bar{x})^2}{(n+1)^2}$$

and $$(x_s - \bar{x}')^2 = \frac{1}{(n+1)^2}((n+1)x_s - n\bar{x} - x_s)^2$$

$$= \frac{n^2}{(n+1)^2}(x_s - \bar{x})^2$$

$$\therefore \Delta C_+ = \frac{n}{(n+1)}(x_s - \bar{x})^2$$

Similarly the change in the contribution to C' of a cluster from which an element is removed is

$$\Delta C = - \frac{n}{(n-1)} (x_s - \bar{x})^2$$

The criterion C' and hence C will be increased by moving $x_s$ from cluster $\ell$ to cluster m if and only if

$$\frac{n_m}{n_m+1} \sum_{j=1}^{J} (x_{s_j} - \bar{x}_{m_j})^2 < \frac{n_\ell}{n_\ell-1} \sum_{j=1}^{J} (x_{s_j} - \bar{x}_{\ell_j})^2$$

Notice that testing for this condition requires only the position of the point in question and those of the two current cluster centres.

## REFERENCES


AUGUSTSON, J. G. & MINKERS, J. (1970). Deriving term relations for a corpus by graph theoretical clusters. Journal of the American Society for Information Science, 1970, 21(2), 101-111.

BATH UNIVERSITY and POLYTECHNIC OF NORTH LONDON. The collection and construction of a file of citation data. By Maureen Ritchie and David Nicholas, Polytechnic of North London and Bath University, 1976. (DISISS Research Report B5)

BATH UNIVERSITY The structure of social science literature as shown by citations. Bath, University Library, 1977 (DISISS Research Report A3).

BATH UNIVERSITY and POLYTECHNIC OF NORTH LONDON. Literature usage and interrelationship in the social sciences as shown by citations in monographs. Bath, University Library, 1977 (DISISS Research Report B7).

BORKO, H. (1965). A factor analytically derived classification system for psychological reports. Perceptual and Motor Skills, 1965, 29, 393-406.

BORKO, H. & BERNICK, M. D. (1963). Automatic document classifications. Journal of the Association for Computing Machinery, 1963, 10, 151-162.

BORKO, H. & BERNICK, M. D. (1964). Automatic document classification part II: additional experiments. Journal of the Association for Computing Machinery, 1964, 11, 138-151.

CARPENTER, M. P. & NARIN F. (1972). Clustering of scientific journals. Chicago, Computer Horizons Inc., 1972.

CUNNINGHAM, K. M. & OCILVIE, J. C. (1972). Evaluation of hierarchical grouping techniques: preliminary study. Computer Journal, 1972, 15(3), 209-213.

DALE, A. G. & DALE, N. (1965). Some clumping experiments for associative document retrieval. American Documentation, 1965, 16(1), 5-9.

DESIGN OF INFORMATION SYSTEMS IN THE SOCIAL SCIENCES (1972). Citation patterns in the social sciences: results of pilot citation study and selection of course journals for main citation study. Bath, Bath University Library, October 1972. (DISISS Working Paper No. 5).

DOYLE, L. B. (1962). Indexing and abstracting by association. American Documentation, 1962, 13(4), 378-390.

DOYLE, L. B. (1968). Some compromises between word grouping and document grouping. Santa Monica, Calif., System Development Corporation, 1964. AD-440 044.

FRIEDMAN, H. P. & RUBIN, J. (1967). On some invariant criteria for groupings data. J. American Statistical Association, 1967, 62(4), 1159-1177.

GITMAN, I (1972). A parameter-free clustering model. Pattern Recognition, 1972, 4(4), 307-315.

GOTTLIEB, C. C. & KUMAR, S. (1968). Semantic clustering of index terms. Journal of the Association for Computing Machinery, 1968, 15(4), 493-513.

HARRISON, P. J. (1968). A method of cluster analysis and some applications. Applied Statistics, 1967, 16(3), 226-236.

HUBERT, L. (1972). Some extensions of Johnson's hierarchical clustering algorithms. Psychometrika, 1972, 37(3), 261-274.

JARDINE, N. (1970). Algorithms, methods and models in the simplification of complex data. Computer Journal, 1970, 13(1), 116-117.

JARDINE, N. & SIBSON, R. (1968). The construction of hierarchic and non-hierarchic classifications. Computer Journal, 1968, 11(2), 117-184.

JOHNSON, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 1967, 32(3), 241-154.

KESSLER, M. M. (1963a). An experimental study of bibliographic coupling between technical papers. IEEE Transactions on Information Theory, 1963, 9(1), 49-51.

KESSLER, M. M. (1963b). Bibliogrphic coupling between scientific papers. American Documentation, 1963, 14(1), 10-25.

KESSLER, M. M. (1965). Comparison of the results of bibliographic coupling with analytic subject indexing. American Documentation, 1965, 16(4), 223-233.

LING, R. F. (1972). On the theory and construction of K-clusters. Computer Journal. 1972, 15(4), 326-332. 

NARIN, F., CARPENTER, M. & BERLT, N. C. (1972). Interrelationships of scientific journals. Journal of the American Society for Information Science, 1972, 23(5), 232-331.

PARKER, E. B., PAISLEY, W. J. & GARRETT, R. (1967). Bibliographic citations as unobtrusive measures of scientific communication. Stanford, Stanford University Institute for Communication Research, 1967.

PREPARATA, F. P. & CHIEN, R. T. (1967). On clustering techniques of citation graphs. Urbana, Ill., University of Illinois, 1967. (Coordinated Science Laboratory Report R-349).

PRICE, N. & SCHIMONOVICH, S. (1968). A clustering experiment: first step towards a computer generated classification scheme. Information Storage and Retrieval, 1968, 4(3), 271-180.

RUBIN, J. (1967). Optimal classification into groups: an approach for solving the taxonomy problem. Theoretical Biology, 1967, 15, 103-144.

SALTON, G. (1965). Process in automatic information retrieval. IEEE Spectrum, 1965, 2,90-103.

SALTON, G. (1969). Information science in a Ph.D. computer science program. Communications of the Association for Computing Machinery, 1969, 12(2), 111-117.

SCHIMINOVICH, S. (1971). Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm. Information Storage and Retrieval, 1971, 6(6), 417-435.

SCOTT, A. J. & SYMONS, M. J. (1971). Clustering methods based on likelihood ratio criteria. Biometrics, 1971, 27(2), 387-397.

SHEPARD, R. N., (1962). The analysis of proximities: multidimensional scaling with an unknown distance function, I and II. Psychometrika, 1962, 27(2), 125-140 and 219-246.

SIBSON, R. (1973). SLINK: an optimally efficient algorithm for the single-cluster method. Computer Journal, 1973, 16, 30-34.

SPARCK JONES, K. (1971). Automatic keyword classification for information retrieval. London, Butterworths, 1971.

STEVENS M. F. (1965). Automatic indexing: a state of the art report. Washington, U.S. Department of Commerce, 1965. (National Bureau of Standards Monograph 91.)

STEVENS, M. F., GIULIANO, V. F. & HEILPRIN, L. B. eds. (1965). Statistical association methods for mechanized documentation. Washington, U.S. Department of Commerce, 1965. (National Bureau of Standards, Miscellaneous Publication 269).

STILES, H. F. (1961). The association factor in information retrieval. Journal of the Association for Computing Machinery, 1961, 8, 271-279.

VAN RIJSBERGEN, C. J. (1970). A clustering algorithm. Computer Journal, 1970, 13(1), 113-115.

WILLIAMS, W. T., LANCE, G. N., DALE, M. B. and CLIFFORD, H. T., (1971). Constroversy concerning the criteria for taxonomic strategies. Computer Journal,1971, 14(2), 162-175.

WOLFF-TERROINE, M. & RIMBERT, D. (1971). Computer aided automatic generation of a structured documentary language: preliminary study. Journal of Documentation, 1971, 127(2), 111-124.

WORONA, S. (1969). Query clustering in a large document space. In: SALTON, G. (1969). Information storage and retrieval. (Section XV). Ithaca, N.Y., Cornell University, Department of Computer Science, 1969.